Wikipedia in the Era of LLMs: Evolution and Risks

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we present a comprehensive analysis and monitoring framework for the impact of Large Language Models (LLMs) on Wikipedia, examining the evolution of Wikipedia through existing data and using simulations to explore potential risks. We begin by analyzing article content and page views to study the recent changes in Wikipedia and assess the impact of LLMs. Subsequently, we evaluate how LLMs affect various Natural Language Processing (NLP) tasks related to Wikipedia, including machine translation and retrieval-augmented generation (RAG). Our findings and simulation results reveal that Wikipedia articles have been affected by LLMs, with an impact of approximately 1%-2% in certain categories. If the machine translation benchmark based on Wikipedia is influenced by LLMs, the scores of the models may become inflated, and the comparative results among models could shift. Moreover, the effectiveness of RAG might decrease if the knowledge has been contaminated by LLMs. While LLMs have not yet fully changed Wikipedia's language and knowledge structures, we believe that our empirical findings signal the need for careful consideration of potential future risks in NLP research.[1]

1 Introduction

The creation of Wikipedia challenged traditional encyclopedias (Giles, 2005), and the rapid development and wide adoption of Large Language Models (LLMs) have sparked concerns about the future of Wikipedia (Wagner & Jiang, 2025; Vetter et al., 2025). Researchers have begun examining the influence of LLMs on Wikipedia, and it is unlikely that Wikipedia has remained unaffected. For example, Reeves et al. (2024) analyze Wikipedia user metrics such as page views and edit histories. Meanwhile, Brooks et al. (2024) estimate the proportion of AI-generated content in newly created English Wikipedia articles using Machine-Generated Text (MGT) detectors. Given the richness and significance of Wikipedia, the impact of LLMs on Wikipedia requires a more comprehensive and detailed investigation.

Wikipedia is widely recognized as a valuable resource (Singer et al., 2017), and its content is extensively utilized in AI research, particularly in Natural Language Processing (NLP) tasks (Johnson et al., 2024b). For instance, Wikipedia pages are among the five datasets used to train GPT-3 (Brown et al., 2020). The sentences in the *Flores-101* evaluation benchmark are extracted from English Wikipedia (Goyal et al., 2022). Lewis et al. (2020)'s work on Retrieval-Augmented Generation (RAG) treated Wikipedia as a source of factual knowledge. Therefore, we aim to investigate the influence of LLMs on machine translation and knowledge systems using Wikipedia as a key resource.

Figure 1 illustrates the various tasks and research topics discussed in this paper. Our first objective is to evaluate the direct impact of LLMs on Wikipedia, focusing on *page views*, *word frequency*, and *linguistic style*. Then we explore the indirect effects on the broader NLP community, particularly in relation to *machine translation benchmarks* and *RAG*, both of which rely heavily on Wikipedia content for their corpora. Thus, we are in a better position to observe and assess the evolutions and risks of Wikipedia in the era of LLMs. Our analysis yields a number of significant insights:

- Some Wikipedia articles have been influenced by LLMs, and the overall impact has so far been quite limited.

[1] We will release all the code and data to the public.

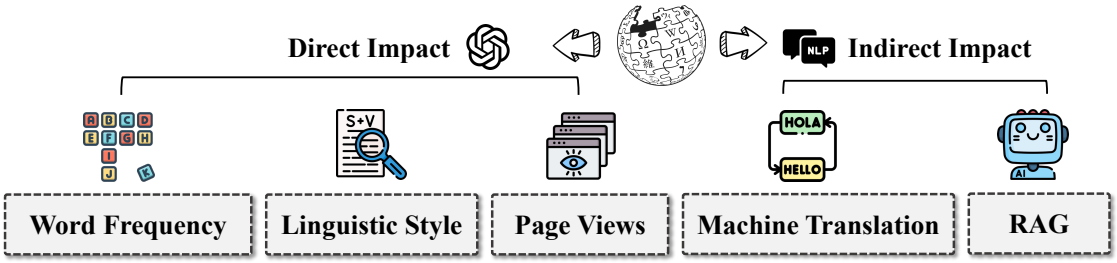


Figure 1: Our work analyze the direct impact of LLMs on Wikipedia, and exploring the indirect impact of LLMs generated content on Wikipedia: **Have LLMs already impacted Wikipedia, and if so, how might they influence the NLP community and human society?**

- There has been a slight decline in page views for certain scientific categories on Wikipedia, although the connection to LLMs remains uncertain.
- If the sentences in machine translation benchmarks are drawn from Wikipedia content shaped by LLMs, the scores of machine translation models are likely to be inflated, potentially reversing the outcomes of comparisons between different models.
- Wikipedia content processed by LLMs could appear less effective for RAG compared to real Wikipedia content.

Based on these findings, we underscore the importance of carefully assessing potential risks and encourage further exploration of these issues. The key contributions of this paper are three-fold, as we are the first to: (1) quantify the impact of LLMs on Wikipedia pages across various categories; (2) analyze the impact of LLMs on Wikipedia based on word usage and provide the corresponding estimates; and (3) examine how LLM-generated content affects machine translation evaluation and the efficiency of RAG systems.

This is also very likely the first paper to comprehensively analyze the impact of LLMs on Wikipedia based on data and simulations. It is important to note that, while some changes are not obvious at the moment, the methods and perspectives we have proposed can be employed for long-term detection of the impact of LLMs on Wikipedia in the future.

2 Related Work

Wikipedia for NLP. Wikipedia has long been utilized in various NLP applications (Strube & Ponzetto, 2006; Mihalcea & Csomai, 2007; Zesch et al., 2008; Gabrilovich & Markovitch, 2009; Navigli & Ponzetto, 2010). In the era of LLMs, Wikipedia also plays an important role, such as in fact-checking (Hou et al., 2024) and reducing hallucinations (Semnani et al., 2023). Writing Wikipedia-like articles is also one of the LLM applications (Shao et al., 2024).

LLMs for Wikipedia. Researchers are trying to use LLMs to enhance Wikipedia, including articles (Adak et al., 2025), Wikidata (Peng et al., 2024; Mihindukulasooriya et al., 2024) and edit process (Johnson et al., 2024a). Some researchers have compared LLM-generated or rewritten Wikipedia articles with human-written ones, yielding differing conclusions Skarlinski et al. (2024); Ashkinaze et al. (2024); Zhang et al. (2025a).

Estimation of LLM Impact. There are studies on the impact of LLMs on its page views (Reeves et al., 2024; Lyu et al., 2025). The detection of AI-generated content has been a hot research topic in recent years (Wu et al., 2025; Wang et al., 2025; Zhang et al., 2024), including its application to Wikipedia articles (Brooks et al., 2024). But MGT detectors have notable limitations (Doughman et al., 2024), and researchers are also exploring other methods for estimating the LLM impact, such as word frequency analysis (Liang et al., 2024; Geng & Trotta, 2024).

3 Data Collection

In this paper, we focus on data from Wikipedia and Wikinews, both under the Wikimedia Foundation.

Wikipedia uses a hierarchical classification system for articles. It begins with top-level categories that cover broad fields, which are then divided into more specific subcategories. Only pages created before 2020 and subcategories that are four or five levels away from our target category were included in our study. Then we scrape the Wikipedia page versions from 2020–2025, using the January 1 snapshot of each year.

We are interested in Wikipedia pages that belong to the following categories: *Art*, *Biology*, *Computer Science (CS)*, *Chemistry*, *Mathematics*, *Philosophy*, *Physics*, *Sports*. Among them, *Philosophy* has the smallest number of articles (33,596), and *CS* leads with the largest number (59,097). More details on data collection and processing are shown in Appendix A.1. For a better comparison, we also collect 6,690 *Featured Articles (FA)*, along with their corresponding 2,029 simple English versions (where available) as *Simple Articles (SA)*.

While Wikipedia is the main focus of this paper, we also collect Wikinews articles from 2020 to 2024 to generate questions in Section 5.2. There are over a hundred news per year, covering a wide variety of topics.

4 Direct Impact from LLMs

4.1 Direct Impact 1: Word Frequency

Since LLMs are extensively applied to writing-related tasks, we aim to investigate whether the text in Wikipedia articles has changed as well. For example, we found that the frequency of certain words favored by LLMs has indeed increased, such as "*crucial*" and "*additionally*" in Figure 2.

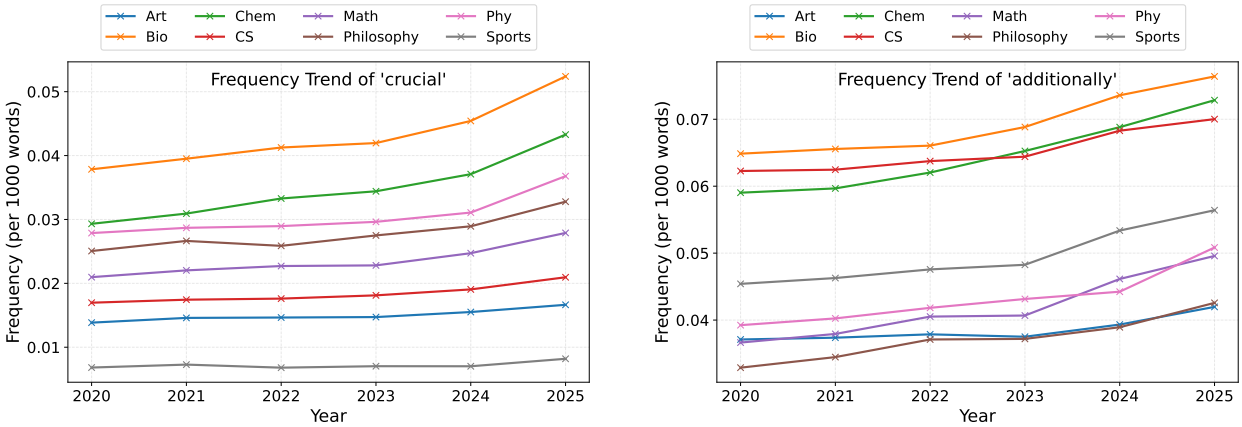


Figure 2: Word frequency in the first section of articles (per 1000 words).

To further investigate whether the changes in word frequency are coincidental or part of a collective shift, we calculate the frequency changes of more words and estimate the impact of LLMs η in one set of Wikipedia articles S based on the following formula (Geng et al., 2025):

$$\hat{\eta}(S) = \frac{\sum_{i \in I} \left(f_i^d(S) - f_i^*(S)\right) f_i^*(S) \hat{r}_i}{\sum_{i \in I} \left(f_i^*(S) \hat{r}_i\right)^2}, \tag{1}$$

$$\hat{r}_i = \frac{f(S_2) - f(S_1)}{f(S_1)}, \tag{2}$$

where $f_i^d(S)$ represents the frequency of word i in the set of texts S, $f_i^*(S)$ represents the one if LLMs do not affect the texts, I is the set of words used for estimation, $f(S_1)$ and $f(S_2)$ represent the frequency of word i for another set of articles before and after LLM processing, respectively.

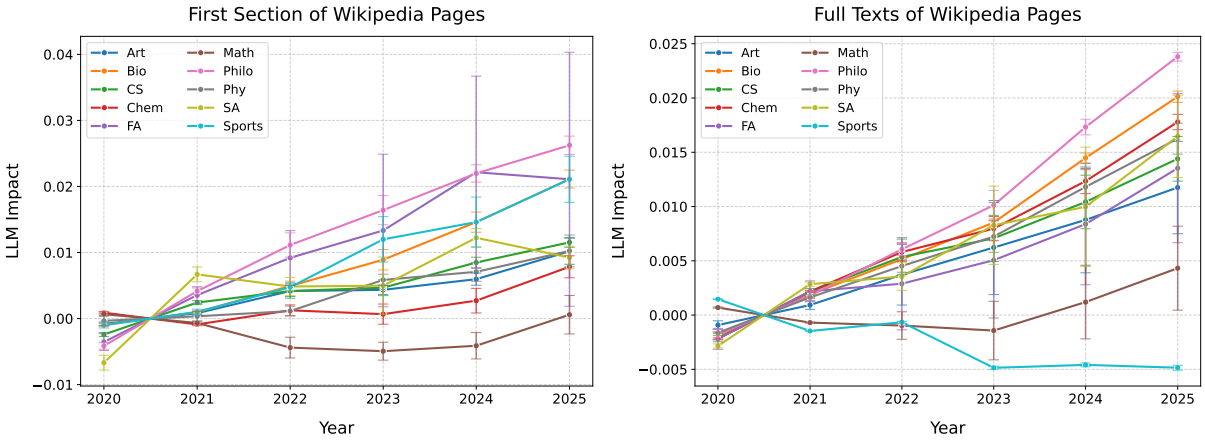


Figure 3: Impact of LLMs on the first section and full content of Wikipedia pages, estimated based on simulations of *Featured Articles*, using different word combinations across different category.

We take the average of the word frequencies from the 2020 and 2021 versions of the page as $f_i^*(S)$. But different texts still lead to different estimations, and using different words for estimation will also produce different results. When estimating r_i through simulations using the first section of *Featured Articles* and *GPT-4o-mini* with a simple prompt: "*Revise the following sentences*", the LLM impact is approximately 1%-2% for the articles in certain categories, as illustrated in Figures 3. Additional results in Appendix A.2 confirm that LLMs have influenced certain categories of Wikipedia articles created before 2020.

Finding 1: While the estimation results vary, the influence of LLMs on Wikipedia is likely to become more significant over time. In some categories, the impact has exceeded 2%.

4.2 Direct Impact 2: Linguistic Style

We also investigate the current and future impact of LLMs on Wikipedia from more linguistic perspectives. In this section, we conduct qualitative rather than quantitative analysis, examining the evolution of Wikipedia content at ***Word***, ***Sentence***, and ***Paragraph*** levels, by comparing the texts before and after LLM processing under the same standards.

Word Level. In addition to word frequency used before, we can also consider it from a broader perspective at the level of words. For instance, the *frequency of auxiliary verbs* indicates the ability of a model to convey complex reasoning and logical relationships (Yang et al., 2024). Lexical diversity, often measured by the *corrected type-token ratio (CTTR)*, reflects the variety of words (Wróblewska et al., 2025). Furthermore, the *proportion of specific parts of speech (POS)* is commonly used as a stylistic feature in assessing the quality of Wikipedia articles (Moás & Lopes, 2023).

Sentence Level. In terms of sentence structure, we focus on *sentence length* and the use of *passive voice* (AlAfnan & MohdZuki, 2023). Regarding sentence complexity, we analyze both the *depth of the entire syntactic tree* and the *clause ratio* (Iavarone et al., 2021).

Paragraph Level. For the paragraph dimension, which is essential for Wikipedia's educational mission (Johnson et al., 2024b), we seek guidance from *readability* evaluation (Moás & Lopes, 2023; Trokhymovych et al., 2024), where six traditional formulas are included in our study: *Automated Readability Index* (Mehta et al., 2018), *Coleman-Liau Index* (Antunes & Lopes, 2019), *Dale-Chall Score* (Patel et al., 2011), *Flesch Reading Ease* (Eleyan et al., 2020), *Flesch–Kincaid Grade Level* (Solnyshkina et al., 2017), and *Gunning Fog index* (Świeczkowski & Kułacz, 2021).

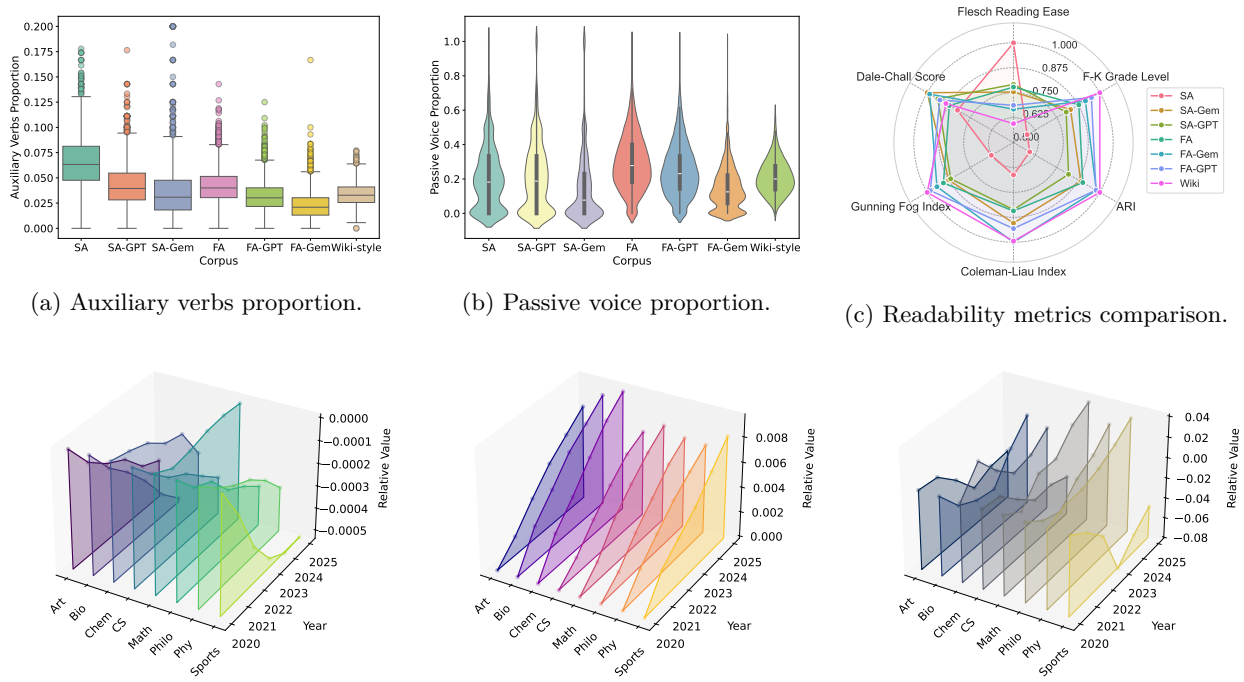
Criteria	LLM-driven trends	Trends in real data	Figures
Auxiliary Verb Rate %	↘	↘	4a, 4d
To be Verb Rate %	↘	↘	12
Corrected Type-Token Ratio (CTTR)	↗	↗	13
Long Words Rate %	↗	↗	14
Conjunction Rate %	–	↗	15a, 15b, 15c
Noun Rate %	↗	↗	15d, 15e, 15f
Preposition Rate %	–	↗	15g, 15h, 15i
Pronouns Rate %	↘	–	15j, 15k, 15l
One-syllable Word Rate %	↘	↘	16a, 16b, 16c
Average Syllables per Word	↗	↗	16d, 16e, 16f
Passive Voice Rate %	↘	↗	4b, 4e
Long Sentence Rate %	↗	↗	17a, 17b, 17c
Average Sentence Length	↗	↗	17d, 17e, 17f
Average Parse Tree Depth	↗	↗	18a, 18b, 18c
Clause Rate %	↗	↗	18d, 18e, 18f
Pronoun-initial Sentence Rate %	↘	↗	19a, 19b, 19c
Article-initial Sentence Rate %	–	↗	19d, 19e, 19f
Dale-Chall Readability	↗	↘	4c, 20a
Automated Readability Index	↗	↗	4c, 20b
Flesch-Kincaid Grade Level	↗	↗	4c, 4f
Flesch Reading Ease	↘	–	4c, 20c
Coleman-Liau Index	↗	–	4c, 20d
Gunning Fox Index	↗	↗	4c, 20e

Table 1: Summary of linguistic style trends. The second column indicates the effects of LLM processing. The third column shows Wikipedia trends over time.

LLM Simulation. Wikipedia articles are not static, and their linguistic styles are difficult to remain the same under different measurement metrics. To understand the link between these trends and LLMs, we simulate the real Wikipedia with *GPT-4o-mini* and *Gemini-1.5-Flash*, then compare the changes before and after the process. Specifically, we instruct both models to revise *Featured* and *Simple* articles using prompt mentioned in Section 4.1, and additionally use *GPT-4o-mini* to generate Wikipedia-style articles using prompt *"Generate a Wikipedia-style article titled {title} and return only the article body in plain text."*

Results. Table 1 presents the summary of the trends in linguistic style in real Wikipedia articles and LLM simulations. The detailed outcomes are illustrated in Figure 4 and Appendix A.3. Although we have plotted the results from 2020 in the these figures, the trends summarized in the table are based on the data in the LLM era, that is, after 2023. For example, our simulation results reveal that LLMs substantially reduce the use of *auxiliary verbs*, with Gemini employing even fewer than GPT, as shown in Figure 4a. Consistent with this tendency, the usage of *auxiliary verbs* on real Wikipedia pages shows a marginal decline from 2020 to 2025, as depicted in Figure 4d. However, the trends of *passive voice proportion* in Figures 4b and 4e are not the same. For *paragraph level*, Figure 4c presents the results of six *readability* metrics, all of which indicate that LLM-generated texts tend to be less readable. The Flesch–Kincaid score in Figure 4f initially decrease and then rises, and the score after LLM simulation also increases.

Finding 2: The trends of changes in Wikipedia articles are largely consistent with the preferences of LLMs under most metrics.



(a) Auxiliary verbs proportion.

(b) Passive voice proportion.

(c) Readability metrics comparison.

(d) Change in auxiliary verbs rate.

(e) Change in passive voice rate.

(f) Change in Flesch–Kincaid.

Figure 4: The results of linguistic style comparison, including the real Wikipedia pages and LLM-simulated pages. The three subplots below represent the differences compared to the data from 2020.

4.3 Direct Impact 3: Page View

The analysis of Wikipedia's page view data can yield many interesting conclusions (Piccardi et al., 2021; 2024). Similar to the work of Reeves et al. (2024), we transform the page view of Wikipedia articles using the inverse hyperbolic sine function. The results are presented in Figures 5 and 21, while the means of the page view values are plotted in Figures 22 and 23 in the appendix.

Notably, there was a slight decline in page views across some scientific categories since 2024. As Reeves et al. (2024) examined changes in Wikipedia page views across various languages from up to January 1, 2024, but our analysis covers pages from different scientific categories and extends to 2025. The latest data actually leads to different findings, and one recent study has reached a similar conclusion (Lyu et al., 2025).

Finding 3: In the second half of 2024, there was a slight decline in page views across some scientific categories, and its connection to the use of LLMs requires further investigation.

5 Indirect Impact from LLMs

5.1 Indirect Impact 1: Machine Translation

Overall. Sentences of some machine translation benchmarks are derived from Wikipedia. If these benchmarks are also influenced by LLMs, what impact would it have on the evaluation results?

Benchmark Construction. We utilize the *Flores dataset*[2], which comprises multiple sentence sets, each representing a single Wikipedia sentence available in several languages. Subsequently, we use *GPT-4o-mini* to translate the English (*EN*) version into the other languages, replacing the original versions to construct the

[2]`https://huggingface.co/datasets/openlanguagedata/flores_plus`

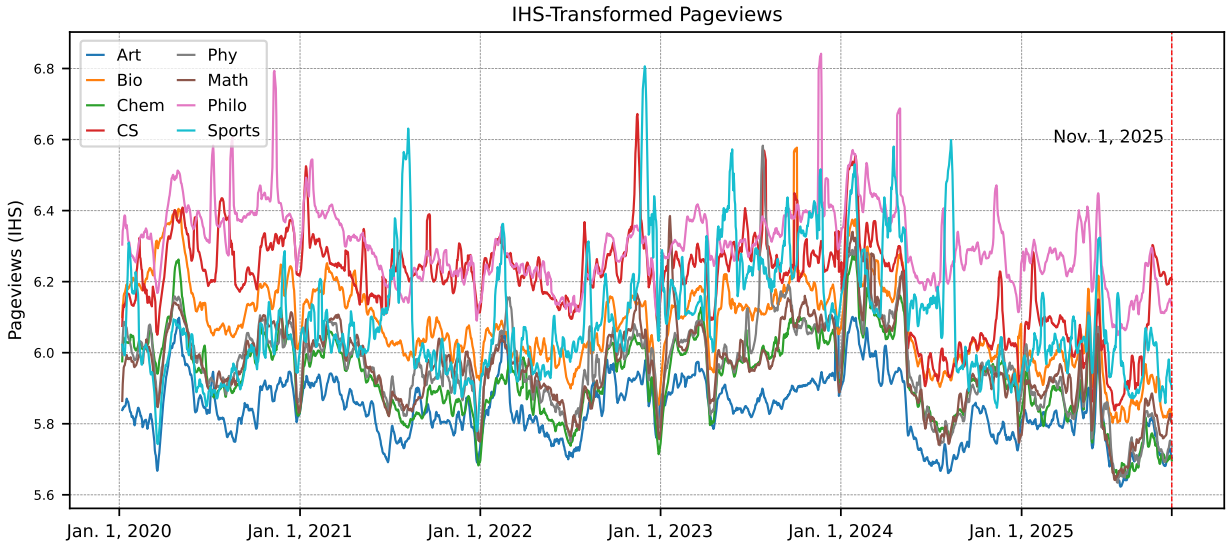


Figure 5: Daily page views across different Wikipedia categories. The y-axis represents page view values after smoothing with a seven-day time window and being transformed via the *Inverse Hyperbolic Sine (IHS)* function.

LLM-influenced benchmark. The following 11 widely used languages are used in our simulations: Modern Standard Arabic (*AR*), Mandarin (*ZH*), German (*DE*), French (*FR*), Hindi (*HI*), Italian (*IT*), Japanese (*JA*), Korean (*KO*), Brazilian Portuguese (*PR*), Russian (*RU*), Latin American Spanish (*ES*).

Evaluation Pipeline. We use different machine translation models to translate machine-translated sentences into other languages, then evaluate them with four metrics: BLEU (Post, 2018), COMET (Rei et al., 2020), ChrF (Popović, 2015), and BERTScore (Zhang et al., 2019).

Models. We compare the translation results from three models: *Facebook-NLLB*[3], a multilingual model supporting 200+ languages (Costa-Jussà et al., 2022); *Google-T5 (mT5)*[4], pre-trained on data covering 101 languages (Xue et al., 2021); and *Helsinki-NLP*'s bilingual Transformer models[5] trained on OPUS corpus (Tiedemann & Thottingal, 2020; Tiedemann et al., 2023).

Results. The results of the comparison between machine translation models could be reversed. For example, *Facebook-NLLB* gets a better BLEU score than *Helsinki-NLP* in the original benchmark, but a worse score in the GPT-processed benchmark, as shown in Tables 2 and 3. In most cases, machine translation models achieve higher scores on the GPT-processed benchmark, as listed in the two tables above and Tables 5, 6, 7, and 8 in the Appendix.

Finding 4: The impact of LLMs on the benchmark could not only inflate the translation scores across different languages but also distort the comparison of translation abilities between models, making it fail to truly reflect their translation effectiveness.

5.2 Indirect Impact 2: RAG

Overall. RAG can provide more reliable and up-to-date external knowledge to mitigate hallucination in LLM generation (Gao et al., 2023). Wikipedia is one of the most commonly applied general retrieval sets in

[3]https://huggingface.co/facebook/nllb-200-3.3B
[4]https://huggingface.co/google/mt5-small
[5]https://github.com/Helsinki-NLP/Opus-MT

	BLEU		ChrF		COMET	
	O	G	O	G	O	G
FR	87.04	96.75	94.62	99.31	90.45	87.79
DE	72.39	93.38	77.98	96.10	84.70	86.37
ZH	72.14	78.61	67.06	78.19	82.40	83.91
AR	71.86	78.73	83.89	88.61	83.19	84.04
PT	69.59	87.71	79.41	92.02	88.93	90.45
JA	62.05	64.21	56.86	58.03	62.61	62.87
ES	59.25	84.44	73.70	90.70	85.03	89.49
IT	58.60	62.14	67.31	78.22	85.22	88.72
HI	58.49	67.29	75.25	80.64	59.53	60.16
KO	54.75	78.35	52.50	69.23	25.94	25.98
RU	51.40	63.33	73.97	84.29	84.75	86.37

Table 2: Facebook-NLLB results on different metrics. *O* and *G* represent the original benchmark and GPT-4o-mini-processed benchmark.

	BLEU		ChrF		COMET	
	O	G	O	G	O	G
FR	88.39	89.40	91.18	91.32	88.39	89.91
DE	68.07	90.68	77.17	94.83	86.35	87.98
ZH	70.34	75.32	59.08	65.10	84.19	85.73
AR	67.52	70.99	80.70	87.20	85.24	86.14
PT	69.74	85.99	81.12	91.60	90.71	92.31
JA	49.48	45.28	49.43	46.40	64.15	64.37
ES	60.00	84.07	74.45	91.26	86.91	91.24
IT	56.14	69.32	67.97	82.04	87.53	90.11
HI	46.85	49.37	58.20	57.06	62.31	63.18
KO	45.28	57.53	58.36	68.94	29.34	29.48
RU	44.99	69.18	70.15	81.81	86.12	87.83

Table 3: Helsinki-NLP results on different metrics. *O* and *G* represent the original benchmark and GPT-4o-mini-processed benchmark.

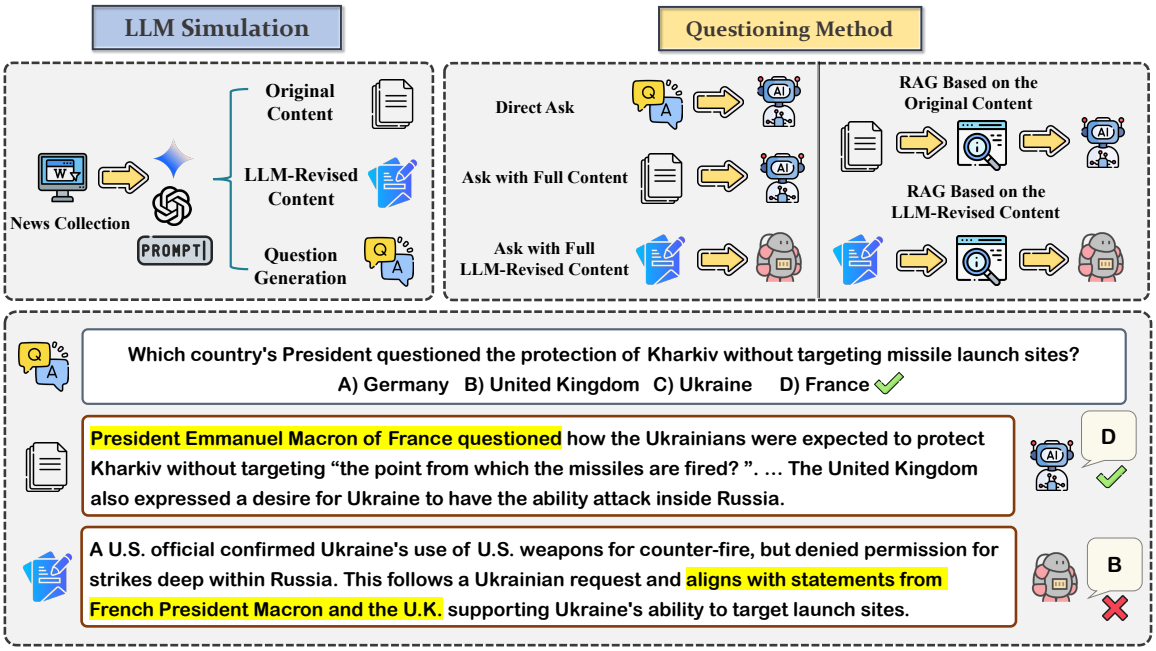


Figure 6: *GPT-4o-mini* and *Gemini-1.5-flash* are used to generate multiple-choice questions (MCQs) based on the extracted Wikinews data. Various questioning methods are employed with *GPT-4o-mini*, *GPT-3.5*, and *DeepSeek-V3* to evaluate the specific impact of LLM-generated texts on the RAG process.

previous RAG work, which stores factual structured information in scale (Fan et al., 2024). In the process of translation using LLMs, some information may also be lost or distorted (Mohamed et al., 2025). Therefore, we are curious how the effectiveness of RAG might change if Wikipedia pages are influenced by LLMs. Our experiment procedure is illustrated in Figure 6 and the detailed steps are listed below.

Question Generation. *GPT-4o-mini* and *Gemini-1.5-flash* are used to generate multiple-choice questions (MCQs) based on Wikinews articles. In order to generate some questions that are not too easy for LLMs, we refer to the prompt in the work of Zhang et al. (2025b), shown in Figure 7.

Prompt

You are to generate three self-contained multiple-choice questions based on the facts mentioned in the following content. Avoid questions that reference the content directly. Each question should include all relevant context and directly name any referenced items, avoiding pronouns like *"it," "the game,"* or *"the person."* Do not include phrases that reference the source or context, such as *"mentioned in the article"* or *"according to the text."*

Figure 7: Prompt for Wikinews-based questions.

Knowledge Base. We construct the knowledge base using Wikinews articles from 2020 to 2024. Each article is preprocessed and split into smaller text segments, then vectorized via BERT[6] (Devlin et al., 2019). We then indexed these vectors using FAISS, a library for efficient similarity search and clustering of dense vectors, for efficient retrieval (Douze et al., 2024).

Retrieval and Generation. The questions are vectorized using BERT and a similarity search is conducted with FAISS. The three most relevant segments are retrieved and used as contextual information. These segments are then combined with the question in a prompt template to query the LLM. The final answer is generated based on both the LLM's prior knowledge and the retrieved content.

Questioning Methods. We conduct experiments using different questioning methods, each involving different LLMs. First, we can query the LLMs directly to obtain answers. Second, the Wikinews page used to generate the question is included in the prompt. Finally, RAG can be used to perform searches in the knowledge base. we also explore different cases, including those using the original Wikinews pages and those using pages processed by the LLMs.

Results. Figure 8 illustrates the summary of the accuracy rates of the LLM responses under different scenarios, with more detailed results provided in Appendix A.6. The analysis based on these results leads to the following conclusions:

- **Higher Accuracy with Knowledge Base.** Providing external knowledge greatly improves performance. With a knowledge base, the accuracy of responses often exceeds 80%. This confirms the effectiveness of RAG in enhancing factual accuracy.
- **Maximal Performance with Full Content.** Providing the full news as context yields the highest accuracy, demonstrating the limitations of retrieval-based approaches in selecting the most relevant information. In most cases with *GPT-4o-mini*, the full content approach exceeded 93% accuracy, setting a benchmark for ideal retrieval performance.
- **Impact of LLM-Revised Content.** Compared to the cases using real Wikinews articles, the accuracy of responses based on ChatGPT-processed pages shows little change and responses based on Gemini-processed pages show a clear drop in accuracy. This suggests that Gemini's rewriting may lead to the loss of some key information.
- **Declining Accuracy for Recent Events.** In the absence of RAG, all models exhibit lower accuracy when answering questions derived from recent Wikinews articles (*e.g.*, *GPT-4o-mini* shown in Table 11 of the appendix: 66.67% in 2024, *GPT-3.5*: 61.25% in 2024), while their accuracy is much better for older events (*e.g.*, 2020–2022). The reason is also straightforward: these news events are not included in their training data. Moreover, DeepSeek-V3 achieves the highest accuracy, which may be attributed to its later knowledge cutoff date.

[6] `https://huggingface.co/bert-base-uncased`

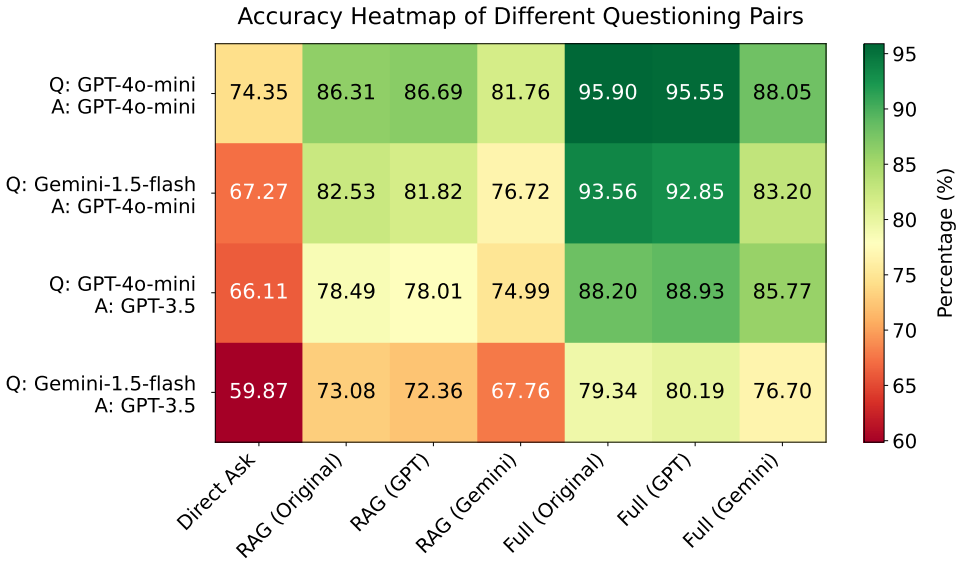


Figure 8: The accuracy rate of LLM responses under different settings. For each case, more than 1,800 questions based on Wikinews articles from 2020 to 2024 are used for simulations. More detailed results are presented in Appendix A.6.

Case Study. To explore the impact of LLM-generated texts, we focus on cases in which the model answers correctly with the original content but fails with the LLM-revised version. Figure 6 has provided one interesting example. The original passage[7] contained an unambiguous clue, "President Emmanuel Macron of France questioned how the Ukrainians were expected to protect Kharkiv ...", which directly supports the correct answer "France." However, in the LLM-revised version, the model reformulated the information into a more abstract and compressed form: "...aligns with statements from French President Macron and the U.K. ..." This revision removes the explicit verb "questioned", merges multiple entities, and relocates key details. As a result, RAG systems relying on the revised text may incorrectly associate the query with "the U.K." due to lexical proximity. This illustrates how LLM-style rewriting can distort relational information and impair factual grounding in RAG systems. More examples are included in Appendix A.6.3, and LLM-generated texts may decrease accuracy in RAG tasks for several reasons:

- **Information Fusion Misleading**: When LLMs merge multiple distinct and clear pieces of information into a single sentence, it can lead to misinterpretation as shown in Figure 6.
- **Keyword Replacement and Omission**: LLM might replace or omit key terms, altering the original meaning and causing misinterpretation in Figures 27, 28 and 29.
- **Abbreviation Ambiguity Misleading**: LLMs use abbreviations inappropriately, leading to misinterpretation as shown in Figure 30.
- **Introduction of Modifiers**: Adding adjectives or modifiers can change the context and impact the text's accuracy, as illustrated in Figure 31.
- **Retrieval Mismatch**: Revised texts may either reduce the similarity between the question and the correct news or increase the similarity with irrelevant ones. In some cases, even small edits to the article lead to a failure in matching.

[7] https://en.wikinews.org/wiki/Ukraine_permitted_to_strike_Russian_territory_near_Kharkiv

Finding 5: The results suggest that LLM-processed content could perform less effectively in RAG systems compared to human-created texts. If such content has impacted high-quality communities like Wikipedia, it raises concerns about the potential decline in information quality in knowledge bases.

6 Discussion and Conclusion

The relationship between Wikipedia and LLMs is bidirectional. On the one hand, Wikipedia content has been a key factor in the growth of LLMs. On the other hand, researchers have used NLP methods, including LLMs, to improve Wikipedia (Lucie-Aimée et al., 2024). Humans and LLMs are coevolving (Geng & Trotta, 2025), and Wikipedia may be one of the bridges in this process. Our study also provides new insights into the risks associated with work that uses Wikipedia data.

In this paper, we collect a large amount of real-world data and conduct comprehensive experimental simulations. Our findings suggest that LLMs are impacting Wikipedia and the impact could extend indirectly to some NLP tasks through their dependence on Wikipedia content. For instance, the target language for machine translation may gradually shift towards the language style of LLMs, albeit in small steps. In addition, the accuracy of RAG tasks may decline when LLM-revised Wikipedia pages are used, indicating potential risks of using LLMs to support Wikipedia or similar knowledge systems.

Although some of the changes may not be immediately apparent, our work offers a framework for extended monitoring longer-term monitoring. These results will also serve as excellent illustrations of the impact of AI on society, given the significant amount of human engagement with Wikipedia. This kind of social impact is already taking place, but has not been adequately addressed by the AI community.

Limitations

Although we conduct several experiments to evaluate the impact of LLMs on Wikipedia, our study has certain limitations. First, some analyses are primarily correlational, identifying patterns but not definitively attributing observed changes to LLMs.

Second, the lack of field experiments limits our insights into the actual machine-in-the-loop editing processes behind Wikipedia article creation. Real-world editing involves complex interactions between humans and sophisticated LLM-based tools. These dynamics may not be fully captured by our simulated studies.

Additionally, when assessing the readability of Wikipedia pages, we rely only on traditional metrics based on formulas, such as the Flesch-Kincaid score. However, recent advances in NLP have shifted towards computational models (François, 2015). Moreover, in the RAG task, our Wikinews dataset is not large enough compared to the Wikipedia page dataset, which may limit the generalization of our findings.

References

Sayantan Adak, Pauras Mangesh Meher, Paramita Das, and Animesh Mukherjee. Reversum: A multi-staged retrieval-augmented generation method to enhance wikipedia tail biographies through personal narratives. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 732–750, 2025.

Mohammad Awad AlAfnan and Siti Fatimah MohdZuki. Do artificial intelligence chatbots have a writing style? an investigation into the stylistic features of chatgpt-4. *Journal of Artificial intelligence and technology*, 3:85–94, 2023.

Hélder Antunes and Carla Teixeira Lopes. Analyzing the adequacy of readability indicators to a non-english language. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pp. 149–155. Springer, 2019.

Joshua Ashkinaze, Ruijia Guan, Laura Kurek, Eytan Adar, Ceren Budak, and Eric Gilbert. Seeing like an ai: How llms apply (and misapply) wikipedia neutrality norms. *arXiv preprint arXiv:2407.04183*, 2024.

Creston Brooks, Samuel Eggert, and Denis Peskoff. The rise of ai-generated content in wikipedia. *arXiv preprint arXiv:2410.08044*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. Exploring the limitations of detecting machine-generated text. *arXiv preprint arXiv:2406.11073*, 2024.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Derar Eleyan, Abed Othman, and Amna Eleyan. Enhancing software comments readability using flesch reading ease score. *Information*, 11:430, 2020.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, 2024.

Thomas François. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20:79–97, 2015.

Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34:443–498, 2009.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Mingmeng Geng and Roberto Trotta. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*, 2024.

Mingmeng Geng and Roberto Trotta. Human-llm coevolution: Evidence from academic writing. *arXiv preprint arXiv:2502.09606*, 2025.

Mingmeng Geng, Caixi Chen, Yanru Wu, Yao Wan, Pan Zhou, and Dongping Chen. The impact of large language models in academia: from writing to speaking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19303–19319, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.987. URL `https://aclanthology.org/2025.findings-acl.987/`.

Jim Giles. Special report internet encyclopaedias go head to head. *nature*, 438:900–901, 2005.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*, 2024.

Benedetta Iavarone, Dominique Brunato, Felice Dell'Orletta, et al. Sentence complexity in context. In *CMCL 2021-Workshop on Cognitive Modeling and Computational Linguistics, Proceedings*, pp. 186–199. Association for Computational Linguistics (ACL), 2021.

Isaac Johnson, Guosheng Feng, Robert West, et al. Edisum: Summarizing and explaining wikipedia edits at scale. *arXiv preprint arXiv:2404.03428*, 2024a.

Isaac Johnson, Lucie-Aimée Kaffee, and Miriam Redi. Wikimedia data for ai: a review of wikimedia datasets for nlp tasks and ai-assisted editing. *arXiv preprint arXiv:2410.08918*, 2024b.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024.

Lucie Lucie-Aimée, Angela Fan, Tajuddeen Gwadabe, Isaac Johnson, Fabio Petroni, and Daniel Van Strien. Proceedings of the first workshop on advancing natural language processing for wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, 2024.

Liang Lyu, James Siderius, Hannah Li, Daron Acemoglu, Daniel Huttenlocher, and Asuman Ozdaglar. Wikipedia contributions in the wake of chatgpt. *arXiv preprint arXiv:2503.00757*, 2025.

Manish P Mehta, Hasani W Swindell, Robert W Westermann, James T Rosneck, and T Sean Lynch. Assessing the readability of online information about hip arthroscopy. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 34:2142–2149, 2018.

Rada Mihalcea and Andras Csomai. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 233–242, 2007.

Nandana Mihindukulasooriya, Sanju Tiwari, Daniil Dobriy, Finn Årup Nielsen, Tek Raj Chhetri, and Axel Polleres. Scholarly wikidata: Population and exploration of conference data in wikidata using llms. In *International Conference on Knowledge Engineering and Knowledge Management*, pp. 243–259. Springer, 2024.

Pedro Miguel Moás and Carla Teixeira Lopes. Automatic quality assessment of wikipedia articles—a systematic literature review. *ACM Computing Surveys*, 56:1–37, 2023.

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, and Guokan Shang. Llm as a broken telephone: Iterative generation distorts information. *arXiv preprint arXiv:2502.20258*, 2025.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 216–225, 2010.

Priti P Patel, Ian C Hoppe, Naveen K Ahuja, and Frank S Ciminello. Analysis of comprehensibility of patient information regarding complex craniofacial conditions. *Journal of Craniofacial Surgery*, 22:1179–1182, 2011.

Yiwen Peng, Thomas Bonald, and Mehwish Alam. Refining wikidata taxonomy using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 5395–5399, 2024.

Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. On the value of wikipedia as a gateway to the web. In *Proceedings of the Web Conference 2021*, pp. 249–260, 2021.

Tiziano Piccardi, Martin Gerlach, and Robert West. Curious rhythms: Temporal regularities of wikipedia consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pp. 1249–1261, 2024.

Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pp. 392–395, 2015.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

Neal Reeves, Wenjie Yin, Elena Simperl, and Miriam Redi. " the death of wikipedia?"–exploring the impact of chatgpt on wikipedia engagement. *arXiv preprint arXiv:2405.10205*, 2024.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.

Sina J Semnani, Violet Z Yao, Heidi C Zhang, and Monica S Lam. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. *arXiv preprint arXiv:2305.14292*, 2023.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*, 2024.

Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read wikipedia. In *Proceedings of the 26th international conference on world wide web*, pp. 1591–1600, 2017.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. Evaluating text complexity and flesch-kincaid grade level. *Journal of social studies education research*, 8:238–248, 2017.

Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pp. 1419–1424, 2006.

Damian Świeczkowski and Sławomir Kułacz. The use of the gunning fog index to evaluate the readability of polish and english drug leaflets in the context of health literacy challenges in medical linguistics: An exploratory study. *Cardiology Journal*, 28:627–631, 2021.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, pp. 713–755, 2023. ISSN 1574-0218. doi: 10.1007/s10579-023-09704-w.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. An open multilingual system for scoring readability of wikipedia. *arXiv preprint arXiv:2406.01835*, 2024.

Matthew A Vetter, Jialei Jiang, and Zachary J McDowell. An endangered species: how llms threaten wikipedia's sustainability. *AI & SOCIETY*, pp. 1–14, 2025.

Christian Wagner and Ling Jiang. Death by ai: Will large language models diminish wikipedia? *Journal of the Association for Information Science and Technology*, 2025.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, et al. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. *arXiv preprint arXiv:2501.11012*, 2025.

Anna Wróblewska, Marceli Korbin, Yoed N Kenett, Daniel Dan, Maria Ganzha, and Marcin Paprzycki. Applying text mining to analyze human question asking in creativity research. *arXiv preprint arXiv:2501.02090*, 2025.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pp. 1–66, 2025.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL `https://aclanthology.org/2021.naacl-main.41`.

Qiyuan Yang, Pengda Wang, Luke D Plonsky, Frederick L Oswald, and Hanjie Chen. From babbling to fluency: Evaluating the evolution of language models in terms of human language acquisition. *arXiv preprint arXiv:2410.13259*, 2024.

Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pp. 1646–1652, 2008.

Jiebin Zhang, J Yu Eugene, Qinyu Chen, Chenhao Xiong, Dawei Zhu, Han Qian, Mingbo Song, Weimin Xiong, Xiaoguang Li, Qun Liu, et al. Wikigenbench: Exploring full-length wikipedia generation under real-world scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5191–5210, 2025a.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 409–436, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.29. URL `https://aclanthology.org/2024.findings-naacl.29/`.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yueheng Zhang, Xiaoyuan Liu, Yiyou Sun, Atheer Alharbi, Hend Alzahrani, Basel Alomair, and Dawn Song. Can llms design good questions based on context? *arXiv preprint arXiv:2501.03491*, 2025b.

A Appendix

A.1 Data Collection and Processing

The detailed classification in Wikipedia poses a problem in our data crawling process: When iteratively querying deeper subcategories without limit, the retrieved pages may become less relevant to the original topic (*i.e.*, the root category). To address this issue, we select an appropriate crawl depth for each category to balance the number of pages with their topical relevance, as shown in Table 4.

Category	Art	Bio	Chem	CS	Math	Philo	Phy	Sports
Crawl Depth	4	4	5	5	5	5	5	4
Number of Pages	57,028	44,617	53,282	59,097	47,004	33,596	40,986	53,900

Table 4: Number of Wikipedia articles crawled per category.

We also exclude redirect pages, as they do not contain independent content but link to other target pages. After crawling the pages, we clean the data by extracting the plain text and removing irrelevant sections such as *"References," "See also," "Further reading," "External links," "Notes,"* and *"Footnotes."* To minimize the impact of topic-specific words, only those rank within the top 10,000 in the Google Ngram dataset[8] are included in the calculations. For Wikinews, we use the *TextExtracts extension*[9], which provides an API to retrieve plain-text extracts of page content.

A.2 LLM Impact

We use GPT-4o-mini to revise the January 1, 2022, versions of *Featured Articles* to construct word frequency data reflecting the impact of large language models (LLMs). This choice is based on the assumption that *Featured Articles* are less likely to be affected by LLMs, given their rigorous review processes and ongoing manual maintenance. To reduce errors caused by incomplete data cleaning, we extract only the first section of each *Featured Article* for revision.

We revise the first section of *Featured Articles* and *Simple Articles* with the prompt *"Revise the following sentences:"* and revise the full texts with the prompt *"Revise the following articles:"*. Some responses are filtered due to the prompt triggering Azure OpenAI's content moderation policy, likely because certain Wikipedia pages contain violent content. Therefore, these pages are excluded from our analysis.

Selecting the appropriate word combinations to estimate the impact of LLMs is crucial. On one hand, by setting a threshold for f^*, we ensure that the target vocabulary appears frequently in the corpus. On the other hand, by setting a threshold for $\hat{r}$, we ensure that these words exhibit a significant frequency change after being processed by the LLM.

For the f^* threshold, we propose two strategies: First, the target words should frequently appear in the first section of *Featured Articles*, as we use this part of the articles for LLM refinement when estimating $\hat{r}$; second, the target words should frequently appear in the target corpus. For the first strategy, when calculating the impact of the LLM on different pages, the selected vocabulary combination remains the same. For the second strategy, the influence on pages of different categories will be estimated using the vocabulary combination corresponding to each category.

A.2.1 Featured Articles and Same Words

We use the first section of *Featured Articles* to request revisions from GPT-4o-mini and calculate the estimated change rate for each word. Then, we select words that are frequently used in the *Featured Articles* and show significant changes in frequency after LLM simulation. This approach allows us to apply the same word combinations to estimate Wikipedia pages across different categories. We change the threshold of f^* and $\hat{r}$ to get a more reliable and stable estimation.

- $\frac{1}{f^*}$: 5000, 7000, 9000, 11000, 13000, 15000
- $\hat{r}$: 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20, 0.21 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

For example, when we take $\frac{1}{f^*} < 5000$ and $\frac{\hat{r}+1}{\hat{r}^2} > 0.21$, the words that satisfy the conditions are: 'making', 'end', 'primarily', 'times', 'next', 'remained', 'however', 'placed', 'people', 'much', 're', 'features', 'success',

[8]Google Ngram dataset: `https://www.kaggle.com/datasets/wheelercode/english-word-frequency-list`
[9]TextExtracts extension: `https://www.mediawiki.org/wiki/Extension:TextExtracts#query+extracts`

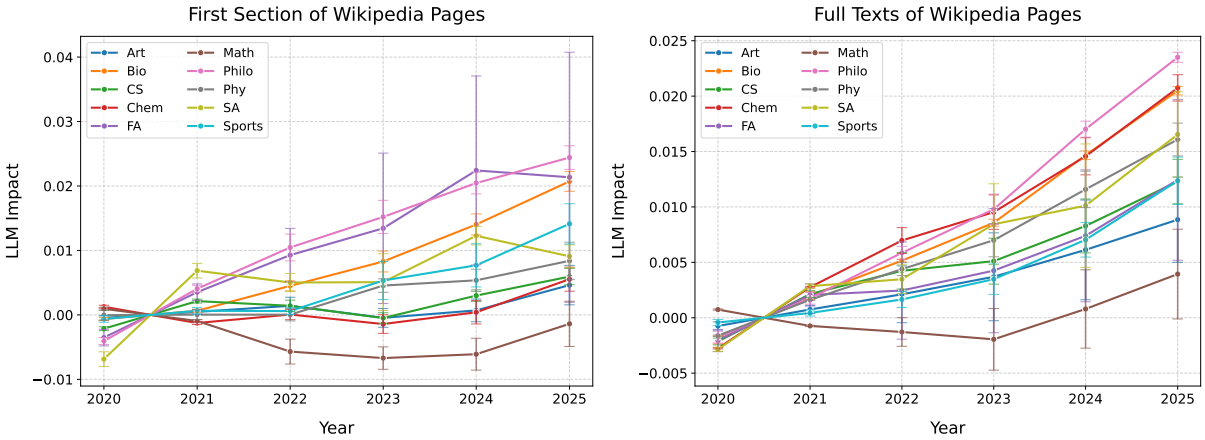


Figure 9: Impact of LLMs on Wikipedia pages, estimated based on simulations of *Featured Articles*, using the same word combinations across each category.

'both', 'down', 'significant', 'appeared', 'formed', 'sent', 'great', 'have', 'numerous', 'but', 'again', 'throughout', 'can', 'country', 'very', 'us', 'book', 'initially', 'based', 'what', 'result', 'because', 'game', 'than', 'remains', 'their', 'once', 'though', 'take', 'described', 'across', 'post', 'went', 'use', 'number', 'successful', 'building', 'win', 'forced', 'run', 'located', 'show', 'combat', 'caused', 'elements', 'victory', 'given', 'today', 'almost', 'while', 'is', 'often', 'following', 'died', 'no', 'make', 'where', 'be', 'popular', 'out', 'upon', 'soon', 'left', 'along', 'wrote', 'total', 'not', 'up', 'were', 'work', 'helped', 'operations', 'written', 'commonly', 'then', 'action', 'long', 'little', 'built', 'worked', 'like', 'created', 'awarded', 'there', 'games', 'although', 'killed', 'attack', 'opened', 'having', 'lived', 'play', 'main', 'few', 'large', 'its', 'important', 'particularly', 'considered', 'p', 'region', 'established', 'coins', 'had', 'major', 'moved', 'more', 'made', 'players', 'these', 'entered', 'spent', 'fought', 'support', 'parts', 'various', 'despite', 'shortly', 'part', 'taken', 'been', 'failed', 'came', 'sometimes', 'launched', 'among', 'during', 'just', 'mostly', 'so', 'this', 'office', 'different', 'player', 'struck', 'forest', 'was', 'called', 'forces', 'would', 'within', 'become', 'story', 'saw', 'last', 'side', 'generally', 'short', 'brought', 'ended', 'won', 'appointed', 'live', 'other', 'best', 'when', 'due', 'introduced', 'largely', 'role', 'men', 'form', 'position', 'served', 'title', 'never', 'including', 'leading', 'way', 'common', 'are', 'man', 'became', 'used', 'about', 'as.'

A.2.2 Featured Articles and Different Words

Unlike the previous strategy which applies the same words across all categories of Wikipedia pages, here we estimate each category using distinct sets of words. For instance, when selecting words for pages in *Computer Science (CS)*, we choose words that frequently appear in *CS* pages and show a relatively higher change rate after LLM simulation. As a result, each category will have its own unique set of words to estimate the impact of LLMs.

- $\frac{1}{f^*}$: 5000, 7000, 9000, 11000, 13000, 15000
- $\hat{r}$: 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20, 0.21 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

For example, when we take $\frac{1}{f^*} < 9000$ and $\frac{\hat{r}+1}{\hat{r}^2} > 0.15$, 635 words in *CS* pages meet these conditions, compared to 496 words in *Art* pages.

A.2.3 Simple Articles and Same Words

The only difference here is that we use *Simple Articles* as the corpus for the LLM simulation process.

- $\frac{1}{f^*}$: 1000, 3000, 5000, 7000, 9000, 11000, 13000
- $\hat{r}$: 0.07, 0.09, 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25, 0.27, 0.29 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

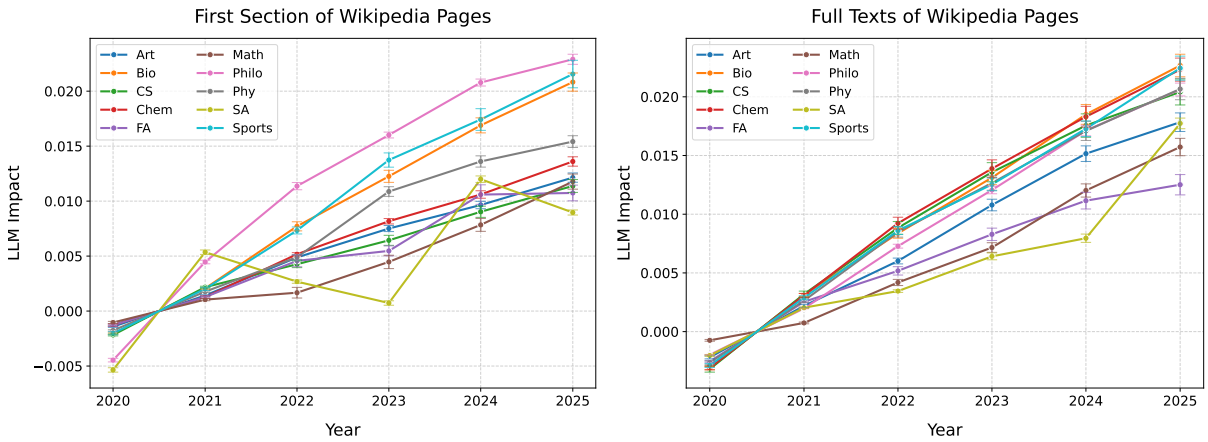


Figure 10: Impact of LLMs on Wikipedia pages, estimated based on simulations of *Simple Articles*, using the same word combinations across each category.

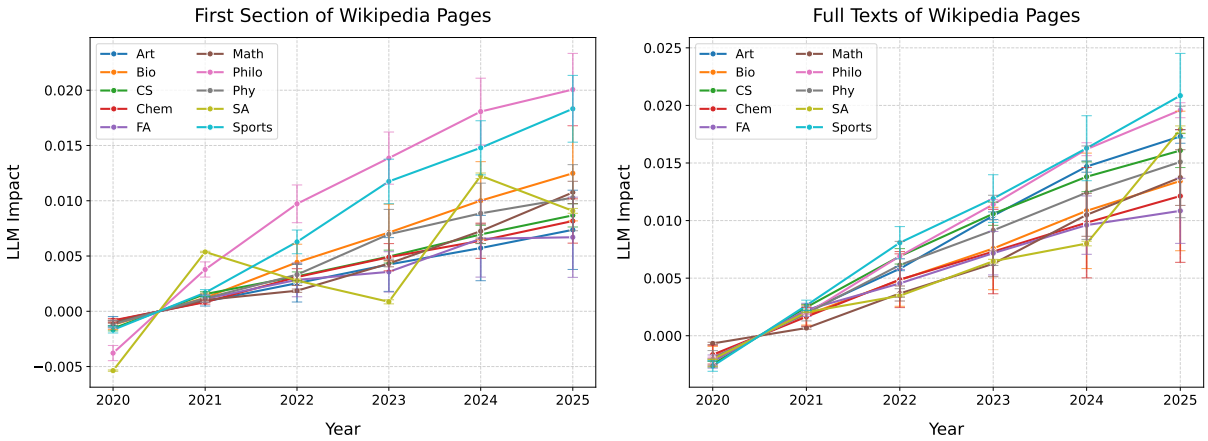


Figure 11: Impact of LLMs on Wikipedia pages, estimated based on simulations of *Simple Articles*, using different word combinations across each category.

A.2.4 Simple Articles and Different Words

- $\frac{1}{f^*}$: 2000, 2500, 3000, 3500, 4000, 4500, 5000
- $\hat{r}$: 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

A.3 Linguistic Style

In this section, we analyze the influence of LLMs on linguistic style among different categories in two dimensions: the first section and full-text content.

A.3.1 Word Level

- ***"To Be"* Verbs** : Figure 12 illustrates that LLMs significantly reduce the usage of *"To Be"* verbs (*e.g.*, replacing *"is important"* with *"demonstrates significance"*), with Gemini using fewer such verbs than GPT. Moreover, a marginal decline in the usage of these verbs is observed in actual Wikipedia pages from 2020 to 2025.
- **Lexical Diversity**: As shown in Figure 13, revised articles display a slightly higher *CTTR*, with texts revised by GPT exhibiting greater lexical diversity than those revised by Gemini. When tasked with

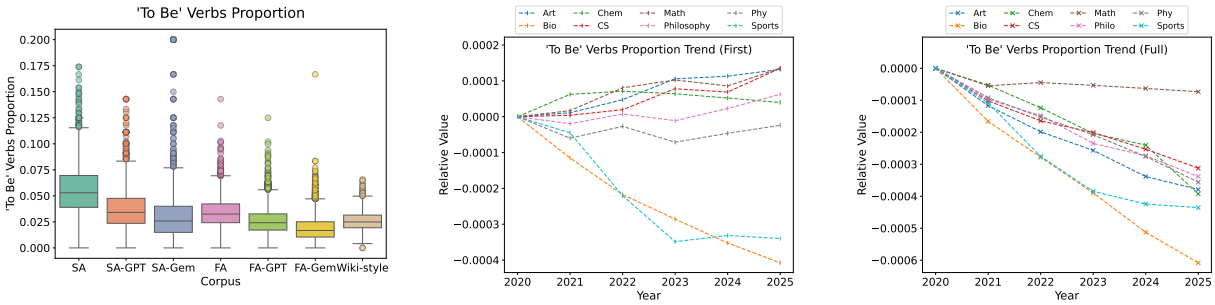


Figure 12: *"To Be"* verbs are reduced by LLMs, with Gemini using fewer than GPT. A slight decline in their usage is also observed in Wikipedia pages from 2020 to 2025.

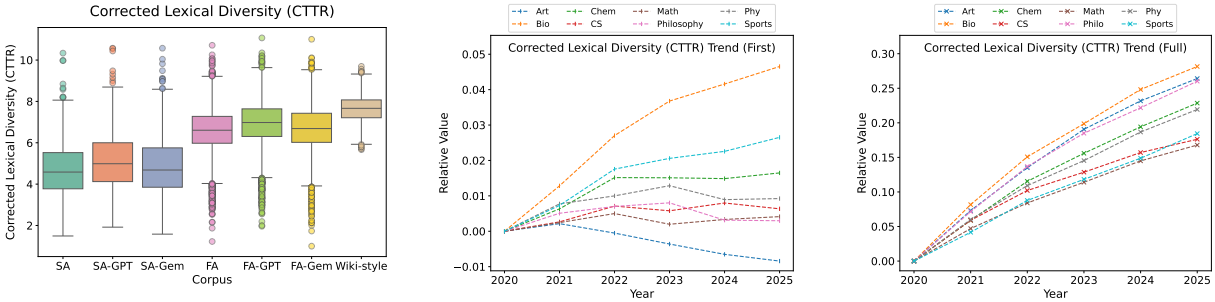


Figure 13: CTTR is slightly higher in revised articles, with GPT showing greater lexical diversity than Gemini. Vocabulary variation has increased across Wikipedia categories over time.

generating wiki-style articles, GPT achieves the highest lexical diversity. Over time, the vocabulary used across different Wikipedia categories has become increasingly varied.

- **Long Words**: Figure 14 indicates that LLMs tend to increase the usage of long words, with Gemini surpassing GPT. From 2020 to 2025, the rate of long words has remained relatively stable across Wikipedia categories.
- **Parts of Speech**: Figure 15 shows that LLMs lead to a slight increase in the use of nouns, accompanied by a corresponding decrease in pronouns. Prepositions and conjunctions remain stable after LLM simulation. On Wikipedia pages, the proportion of prepositions has steadily increased, while the proportions of other parts of speech have remained stable.
- **Syllables**: Figure 16 illustrates that the proportion of one-syllable words declines in articles revised by LLMs, with Gemini employing even fewer such words. Meanwhile, the average syllables per word increase, suggesting a preference for polysyllabic words by LLMs. However, these two metrics remain relatively stable across different Wikipedia categories.

A.3.2 Sentence Level

- **Sentence Length**: Figure 17 shows that both the average sentence length and the proportion of long sentences show a significant increase after being processed by the LLM. Additionally, the period from 2020 to 2025 has seen a notable rise in these two metrics across Wikipedia pages, indicating a trend towards longer sentence structures.
- **Sentence Complexity**: According to figure 18, after revisions by GPT, *Simple Articles* show an increase in complexity, while *Featured Articles* exhibit only minor changes. This may suggest that LLMs do not generate sentences at the highest possible complexity, but instead maintain complexity at a certain level.

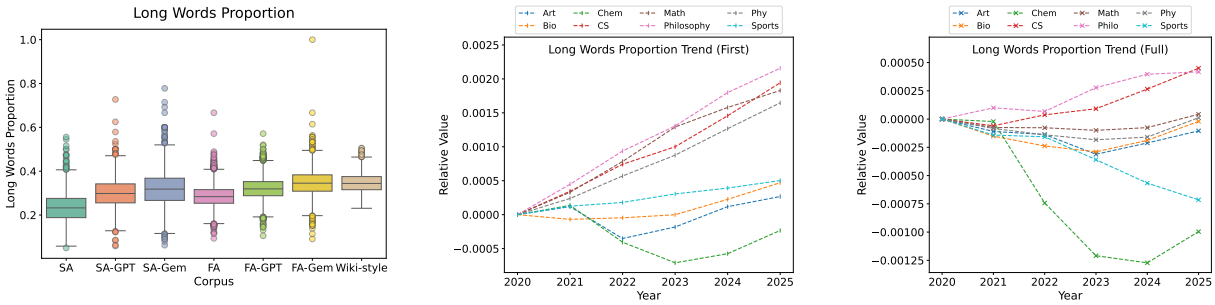


Figure 14: Long words are used more frequently by LLMs, with Gemini surpassing GPT. Their rate has remained stable across Wikipedia categories from 2020 to 2025.

For real Wikipedia pages, a steady year-on-year increase in these two metrics has been observed, indicating a shift towards more complex sentence structures.

- **Pronoun and Article-Initial Sentences**: LLMs tend to avoid starting sentences with pronouns (*e.g.*, "It") or articles (*e.g.*, "The"), as shown in figure 19. For example, it might replace *"The team worked hard to finish the project on time."* with *"Hard work from the team ensured the project was completed on time."* However, in real Wikipedia pages, Article-initial sentences have increased, while pronoun-initial sentences remain stable from 2020 to 2025.

A.3.3 Paragraph Level

We use *Textstat*[10] to calculate six paragraph metrics. *Textstat* is an easy-to-use library to calculate statistics from the text. It provides a range of functions to analyze readability, sentence length, syllable count, and other important textual features.

Through the LLM simulation process, we discover that LLMs tend to generate articles that are harder to read. Figure 20 suggests that the readability of Wikipedia pages has shown only slight variation over the years and does not appear to be influenced by LLMs at this stage.

[10] `https://github.com/textstat/textstat`

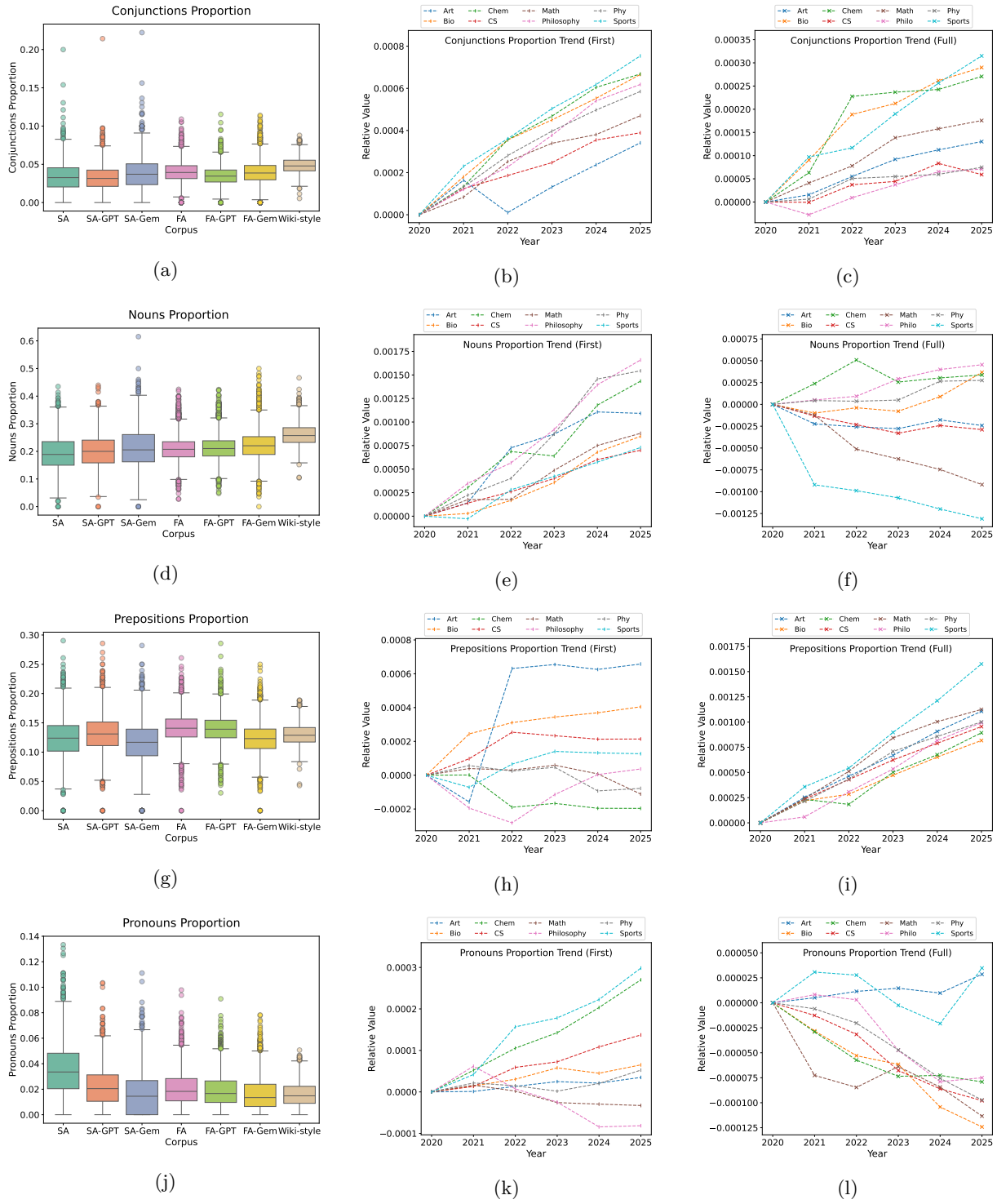


Figure 15: Parts of speech distribution, indicating that LLMs slightly increase nouns and decrease pronouns, while prepositions and conjunctions remain stable. On Wikipedia pages, the proportion of prepositions has steadily increased, with other parts of speech remaining stable.

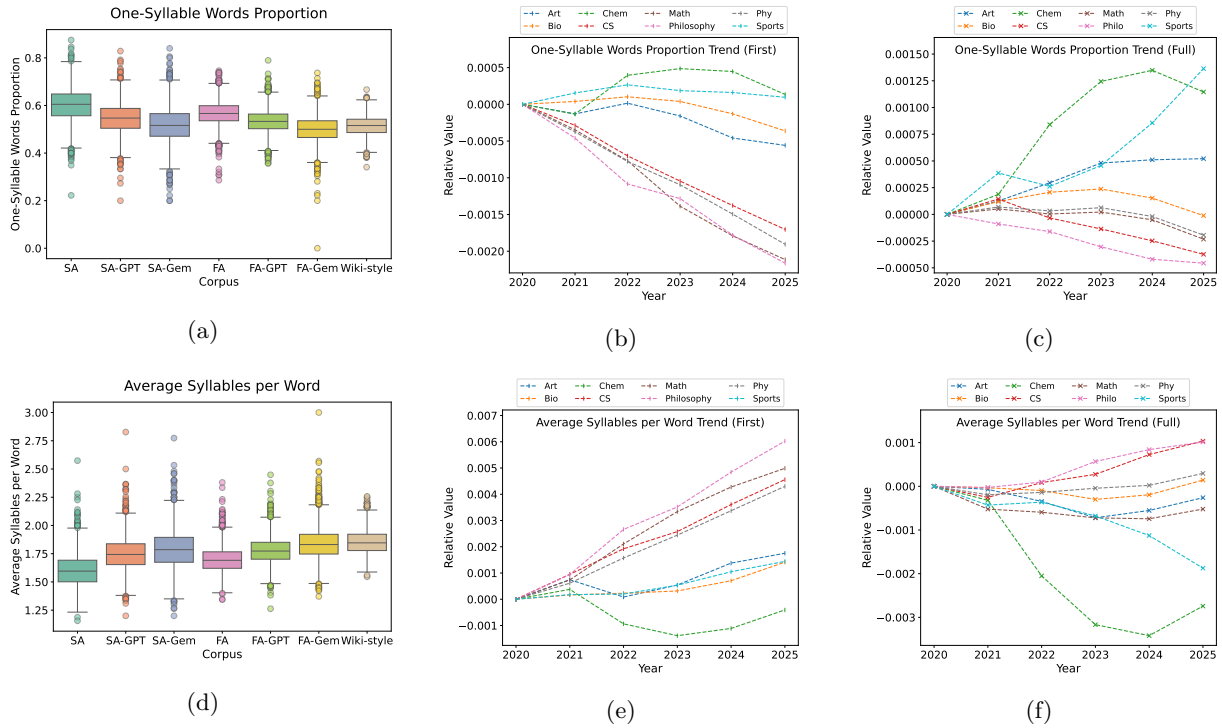


Figure 16: LLMs show a preference for polysyllabic words while reducing the frequency of monosyllabic terms. These two metrics remain relatively stable across different Wikipedia categories.

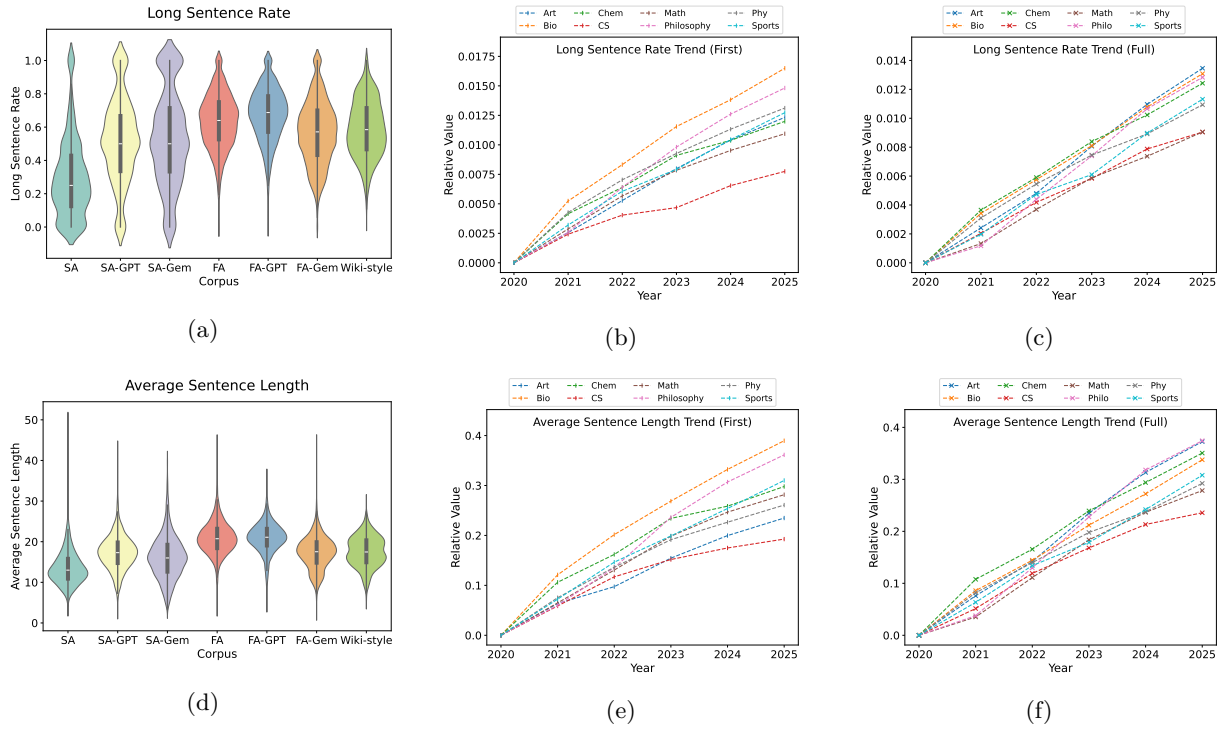


Figure 17: LLMs tend to generate texts with longer sentences, a trend that has grown steadily across Wikipedia categories over the years.

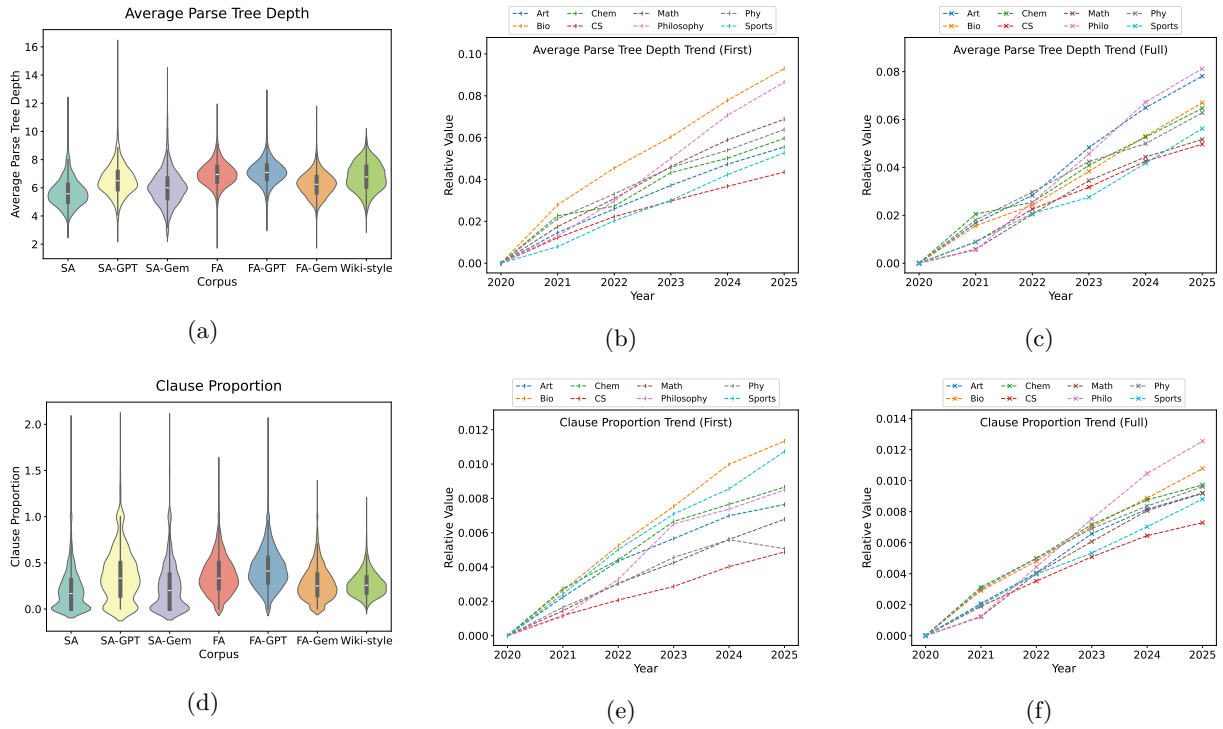


Figure 18: Average parse tree depth and clause proportion remain relatively stable after simulation. In contrast, for actual Wikipedia pages, a gradual year-over-year increase in these two metrics has been observed, indicating a shift towards more complex sentence structures.

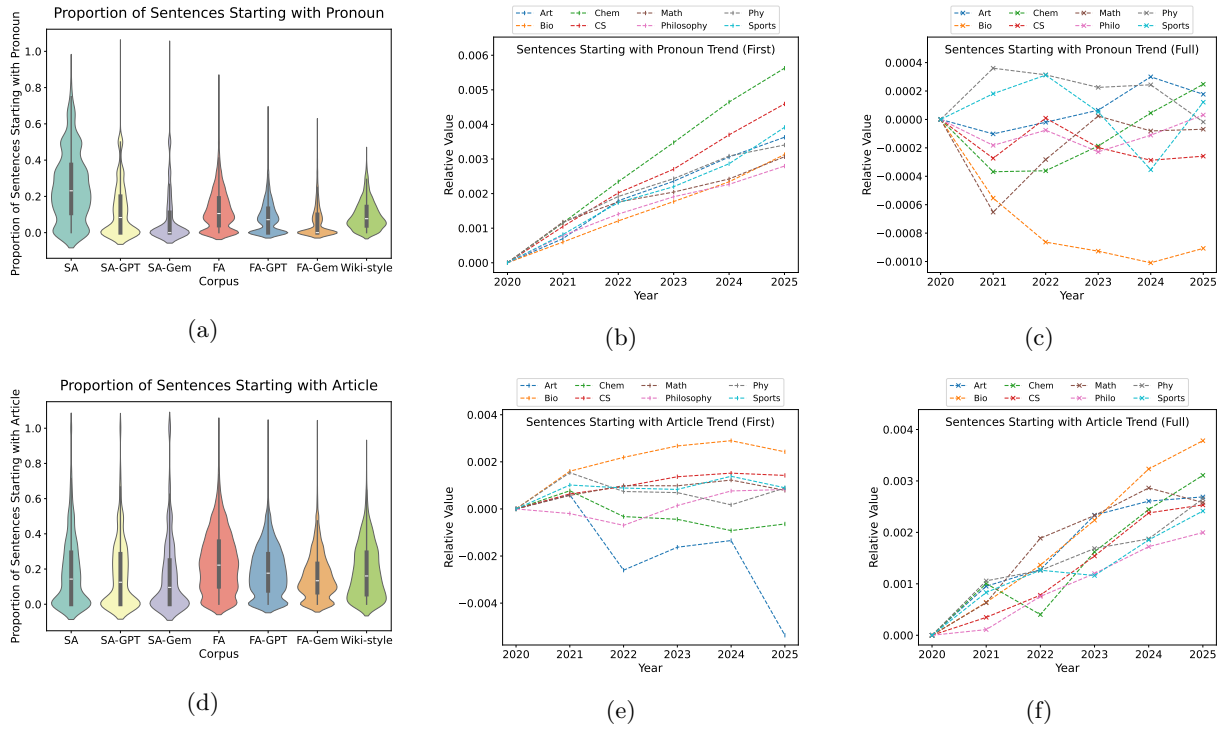
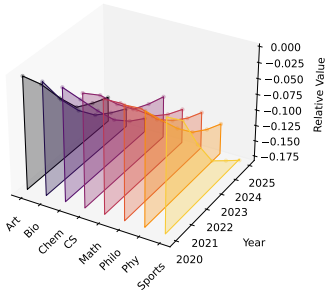
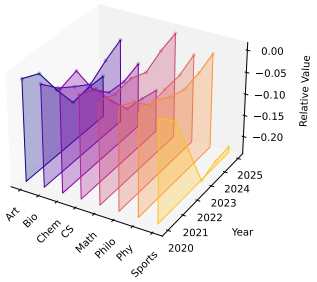


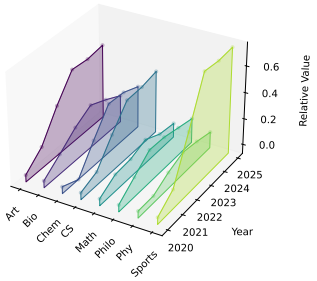
Figure 19: The proportions of sentences starting with specific parts of speech, indicating that LLMs tend to avoid beginning sentences with pronouns or articles.



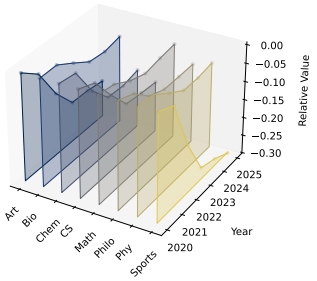
(a) Change in Dale-Chall readability.



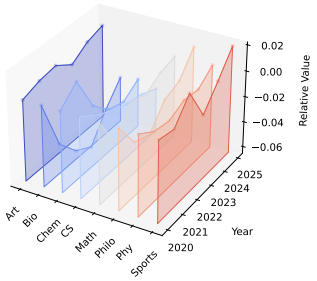
(b) Change in Automated Readability.



(c) Change in Flesch Reading Ease.



(d) Change in Coleman-Liau Index.



(e) Change in Gunning Fog Index.

Figure 20: Changes in readability metrics of Wikipedia pages.

A.4 Page views

The ten categories in our dataset each exhibit unique participation patterns, making comparisons both within and between categories quite challenging. To address this issue, we apply the inverse hyperbolic sine (IHS) function to standardize the page view across different categories. Since Featured articles typically have higher page views and Simple articles lower ones, we present two figures, including and excluding these two categories, to enable fairer comparison. Figure 21 shows the full results with Featured articles and Simple articles included. We also calculate the page views using the *arithmetic mean.* Figures 22 and 23 illustrate the average page views across ten categories. We present an additional result excluding data from *Featured Articles* and *Simple Articles* to better compare other categories. The legends in these three figures are consistent with that of Figure 5.

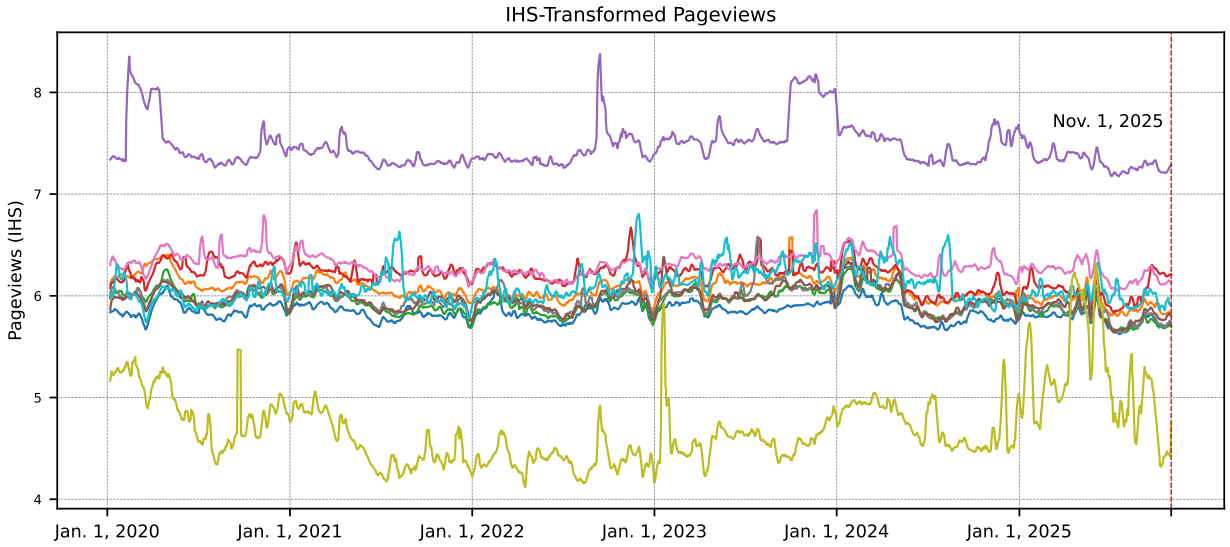


Figure 21: Daily page views across different Wikipedia categories. The y-axis represents page view values transformed via the *Inverse Hyperbolic Sine (IHS)* function.

A.5 Machine Translation

We construct benchmarks influenced by LLMs with the prompt *"Translate the following text to {target language}:"*. Some API calls in our code returned an *openai.BadRequestError* with *error code 400*, indicating that Azure OpenAI's content management policies flagged the prompts for potentially violating content. Also, Some translations returned null values. These cases were excluded from scoring and ignored in the evaluation.

These are the 12 languages in our benchmarks: *English* (eng-Latn-stan1293), *Modern Standard Arabic* (arb-Arab-stan1318), *Mandarin* (cmn-Hans-beij1234), *German* (deu-Latn-stan1295), *French* (fra-Latn-stan1290), *Hindi* (hin-Deva-hind1269), *Italian* (ita-Latn-ital1282), *Japanese* (jpn-Jpan-nucl1643), *Korean* (kor-Hang-kore1280), *Brazilian Portuguese* (por-Latn-braz1246), *Russian* (rus-Cyrl-russ1263), *Latin American Spanish* (spa-Latn-amer1254).

For *Google-T5* shown in Table 5, German (*DE*) initially has a *BLEU* score of 30.24, which rises to 44.18 in the GPT-processed benchmark, marking another substantial improvement.

We also evaluate our results using BERTScore, as shown in Tables 6, 7, and 8. Precision measures how many tokens in the candidate sentence are similar to tokens in the reference sentence, capturing how much of the candidate sentence is relevant to the reference. Recall measures how many tokens in the reference sentence are similar to tokens in the candidate sentence, capturing how much of the reference sentence is represented in the candidate. As for F1 Score, BERTScore combines precision and recall into an F1 score,

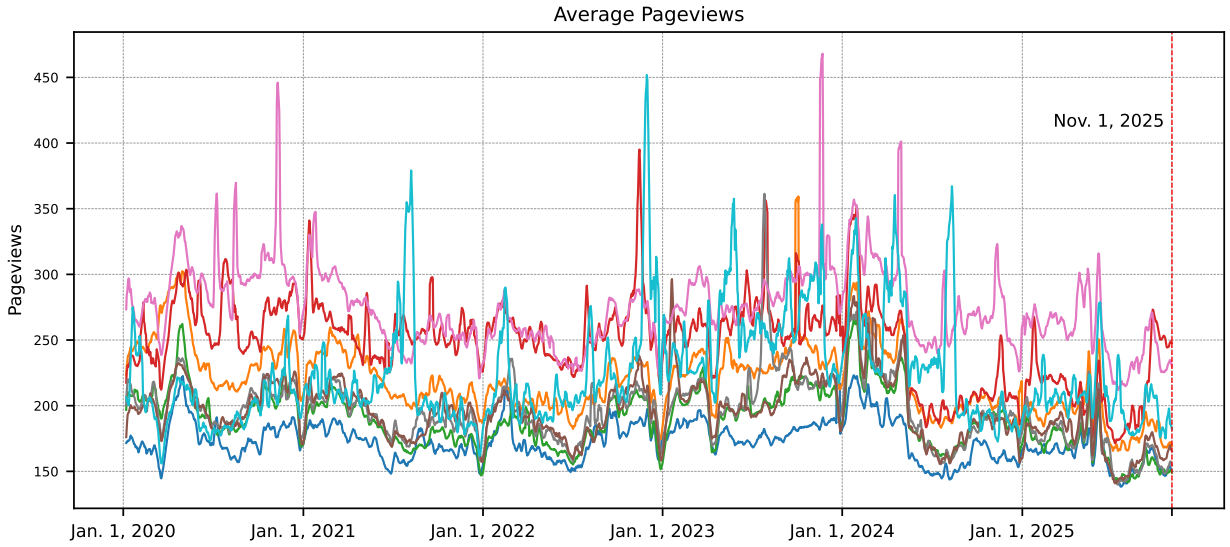


Figure 22: Daily page views across different Wikipedia categories. The y-axis represents page view values transformed via arithmetic mean.

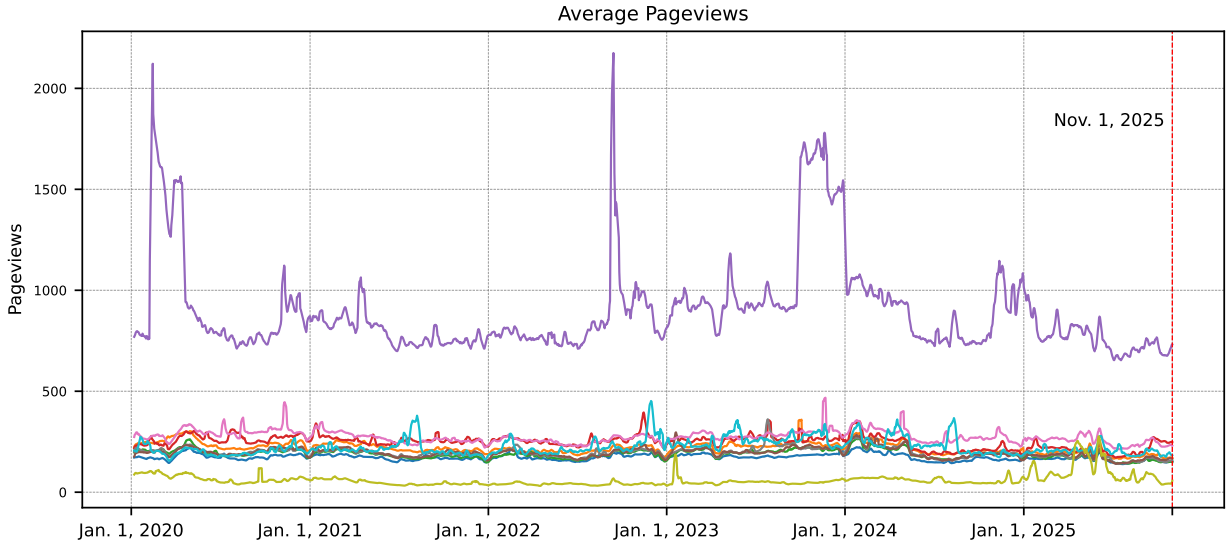


Figure 23: Daily page views across different Wikipedia categories. The y-axis represents page view values transformed via arithmetic mean.

the harmonic mean of the two. This balanced measure provides a single metric that reflects the accuracy and completeness of the candidate sentence relative to the reference.

Overall, our conclusion that LLM-influenced benchmarks inflate translation scores across different languages still holds when using BERTScore as the evaluation metric.

	BLEU		ChrF		COMET	
	O	G	O	G	O	G
DE	71.52	80.09	84.27	93.62	83.91	85.63
FR	68.33	65.93	87.86	86.32	85.49	87.01

Table 5: Google-T5 results on some metrics.

	Precision		Recall		F1	
	O	G	O	G	O	G
AR	0.869	0.891	0.854	0.880	0.861	0.886
DE	0.890	0.915	0.874	0.900	0.882	0.907
ES	0.885	0.953	0.867	0.943	0.876	0.948
FR	0.919	0.944	0.902	0.930	0.910	0.937
HI	0.876	0.900	0.865	0.892	0.870	0.896
IT	0.883	0.936	0.865	0.925	0.874	0.930
JA	0.829	0.850	0.808	0.824	0.818	0.836
KO	0.849	0.878	0.842	0.869	0.845	0.873
PT	0.923	0.946	0.913	0.937	0.918	0.941
RU	0.875	0.908	0.858	0.892	0.866	0.899
ZH	0.839	0.861	0.778	0.797	0.806	0.827

Table 6: BERTScore evaluation results on the Facebook-NLLB translation outputs.

	Precision		Recall		F1	
	O	G	O	G	O	G
AR	0.861	0.872	0.854	0.869	0.857	0.870
DE	0.896	0.923	0.888	0.916	0.892	0.919
ES	0.884	0.952	0.871	0.949	0.877	0.951
FR	0.919	0.944	0.911	0.941	0.915	0.942
HI	0.812	0.822	0.785	0.798	0.798	0.809
IT	0.880	0.935	0.869	0.931	0.874	0.933
JA	0.608	0.612	0.625	0.629	0.617	0.620
KO	0.610	0.614	0.602	0.605	0.605	0.609
PT	0.929	0.954	0.926	0.951	0.927	0.952
RU	0.868	0.896	0.859	0.889	0.863	0.892
ZH	0.852	0.870	0.825	0.845	0.838	0.857

Table 7: BERTScore evaluation results on the Helsinki-NLP machine translation outputs.

	Precision		Recall		F1	
	O	G	O	G	O	G
DE	0.873	0.898	0.845	0.869	0.858	0.883
FR	0.887	0.907	0.849	0.869	0.867	0.887

Table 8: BERTScore evaluation results (Precision, Recall, F1) on the Google-T5 translation outputs.

A.6 RAG

A.6.1 Experiment Setup

Table 9 presents the LLM parameters employed in RAG simulations, such as the *knowledge cutoff date*, *temperature*, and *top-p*. Table 10 shows the annual number of questions generated by different LLMs.

We revise Wikinews articles with the prompt *"Revise the following news."* Figures 24, 25, and 26 present the prompts used for different questioning methods.

A.6.2 Detailed Results

Tables 11, 12, 13, 14, 15, and 16 present RAG results where *"Null"* output is counted as 0 accuracy, while Tables 17, 18, 19, 20, 21, and 22 display results with *"Null"* output counted as 0.25 accuracy. Questions generated from Wikinews in 2024 are likely the most up-to-date; therefore, we focus on their results, which correspond to the last row in each table. Key figures are highlighted in bold. The results indicate that LLM-revised content tends to be less effective as a knowledge source, as accuracy based on LLM-revised texts is often lower than that based on the original texts.

Models	Knowledge Cutoff	Temperature	Top-p
GPT-3.5	September 2021	1.0	1.0
GPT-4o-mini	October 2023	1.0	1.0
Gemini-1.5-flash	May 2024	1.0	0.95
DeepSeek-V3	July 2024	1.0	1.0

Table 9: LLM parameters Used in RAG simulations.

Year	2020	2021	2022	2023	2024
Number of GPT genertated Questions	348	453	390	426	240
Number of Gemini genertated Question	348	453	393	426	240

Table 10: Annual Number of Questions Generated by Different LLMs.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	75.86%	85.34%	85.63%	79.60%	95.98%	95.40%	87.36%
2021	71.74%	86.31%	88.96%	79.69%	96.03%	96.03%	88.08%
2022	80.00%	89.49%	87.18%	84.10%	95.64%	95.64%	88.97%
2023	77.46%	87.09%	87.09%	83.33%	96.01%	94.84%	87.09%
2024	66.67%	83.33%	84.58%	82.08%	95.83%	95.83%	88.75%

Table 11: GPT-4o-mini performance on RAG task (problem generated by GPT).

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	66.95%	82.76%	82.47%	75.86%	93.68%	91.38%	84.20%
2021	64.68%	81.90%	82.34%	75.06%	94.04%	93.82%	82.12%
2022	73.54%	86.01%	85.75%	78.88%	94.66%	93.89%	83.21%
2023	69.95%	82.39%	83.10%	78.40%	92.49%	92.25%	83.57%
2024	61.25%	79.58%	75.42%	75.42%	92.92%	92.92%	82.92%

Table 12: GPT-4o-mini performance on RAG task (problem generated by Gemini).

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	68.68%	77.59%	78.16%	74.14%	86.21%	87.93%	87.36%
2021	67.11%	79.25%	79.25%	74.17%	87.42%	88.30%	84.99%
2022	70.26%	82.82%	80.77%	78.97%	88.46%	90.51%	88.46%
2023	64.08%	74.88%	76.06%	71.83%	86.85%	88.73%	84.27%
2024	60.42%	77.92%	75.83%	75.83%	92.08%	89.17%	83.75%

Table 13: GPT-3.5 Performance on RAG task (problem generated by GPT).

A.6.3 Case Study

Figures 27, 28, 29, 30 and 31 present cases where answers are accurate using the original texts but become inaccurate using LLM-revised texts.

Prompt: Direct Asking

```
prompt = (
    f"ˆAnswer the following questions.  The format should be as per 1.  C)..."
    f"Need to answer all questions and mark the question number."
    f"Only need to give each answer without explanation.  Questions:  {text}"
    f"The format should be as per 1.  C)...\n2.  C)..."
    f"All questions are required to be answered.  Don't skip any."
)
```

Figure 24: Prompt used in the direct asking setting.

Prompt: Asking with a knowledge base

```
prompt = (
    f"Use context to answer user questions."
    f"Question:  {question}\n"
    f"Reference context:  {topk_ans}\n"
    f"Only need to give the correct option without explanation."
)
```

Figure 25: Prompt used in the asking with a knowledge base setting.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	66.95%	72.70%	72.41%	68.97%	77.87%	79.31%	77.59%
2021	58.72%	73.73%	71.74%	68.21%	81.02%	79.47%	74.17%
2022	62.09%	74.05%	72.77%	69.47%	82.44%	82.19%	80.41%
2023	56.57%	73.24%	74.88%	67.14%	77.46%	79.58%	74.65%
2024	55.00%	71.67%	70.00%	65.00%	77.92%	80.42%	76.67%

Table 14: GPT-3.5 Performance on RAG task (problem generated by Gemini).

Prompt: Full Texts Provided

```
prompt = (
    f"Use context to answer user questions."
    f"question:  {question}"
    f"Reference context:  {content}"
    f"Only need to give the correct option without explanation.  Don't miss ')' or
option."
    f"If there is no answer in the content, just return None.  Don't give a
string."
)
```

Figure 26: Prompt used in the full texts provided setting.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	89.66%	81.03%	81.90%	78.45%	98.28%	97.70%	90.52%
2021	84.55%	77.26%	79.25%	69.09%	97.57%	97.79%	87.20%
2022	90.00%	80.77%	81.54%	75.90%	97.69%	97.44%	90.00%
2023	83.57%	73.00%	76.29%	69.72%	96.71%	95.54%	88.03%
2024	82.08%	75.42%	72.50%	75.42%	97.08%	96.67%	86.25%

Table 15: DeepSeek-V3 Performance on RAG task (problem generated by GPT).

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	83.62%	74.14%	75.57%	69.54%	94.54%	95.11%	83.05%
2021	54.97%	69.98%	72.41%	63.36%	95.81%	94.70%	81.68%
2022	84.48%	78.12%	78.37%	65.39%	96.18%	94.91%	84.99%
2023	65.02%	73.00%	74.88%	64.55%	95.54%	94.37%	83.33%
2024	75.83%	74.17%	70.42%	70.00%	95.42%	93.75%	84.58%

Table 16: DeepSeek-V3 Performance on RAG task (problem generated by Gemini).

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	75.86%	85.76%	86.28%	80.03%	96.19%	95.76%	89.15%
2021	71.74%	86.53%	89.24%	80.08%	96.25%	96.36%	89.85%
2022	80.00%	89.87%	88.14%	84.55%	95.90%	95.96%	90.51%
2023	77.52%	87.44%	87.32%	83.69%	96.24%	95.18%	89.14%
2024	67.60%	83.75%	85.21%	82.92%	96.15%	96.15%	90.10%

Table 17: GPT-4o-mini performance on RAG task (problem generated by GPT), Null Output is counted as 0.25.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	67.53%	82.90%	82.54%	76.29%	93.75%	91.45%	85.70%
2021	65.01%	81.95%	82.40%	75.22%	94.21%	93.87%	83.83%
2022	73.98%	86.20%	85.94%	79.07%	94.85%	94.08%	84.80%
2023	70.42%	82.63%	83.39%	78.64%	92.72%	92.55%	85.27%
2024	62.50%	80.00%	75.83%	75.94%	93.65%	93.33%	85.00%

Table 18: GPT-4o-mini performance on RAG task (problem generated by Gemini) , Null Output is counted as 0.25.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	68.68%	77.59%	78.16%	74.14%	86.35%	87.93%	87.36%
2021	67.11%	79.25%	79.25%	74.17%	87.42%	88.30%	85.15%
2022	70.26%	82.82%	80.77%	78.97%	88.59%	90.51%	88.65%
2023	64.08%	74.88%	76.06%	71.83%	86.91%	88.79%	84.51%
2024	60.42%	77.92%	75.83%	75.83%	92.29%	89.17%	83.75%

Table 19: GPT-3.5 performance on RAG task (problem generated by GPT) , Null Output is counted as 0.25.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	66.95%	72.70%	72.49%	68.97%	77.95%	79.31%	77.66%
2021	58.72%	73.79%	71.74%	68.21%	81.13%	79.53%	74.34%
2022	62.28%	74.11%	72.84%	69.53%	82.44%	82.25%	80.47%
2023	56.57%	73.24%	74.88%	67.14%	77.70%	79.69%	74.82%
2024	55.00%	71.67%	70.00%	65.00%	78.12%	80.52%	76.67%

Table 20: GPT-3.5 performance on RAG task (problem generated by Gemini) , Null Output is counted as 0.25.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	89.66%	84.48%	85.20%	82.18%	98.42%	97.99%	92.03%
2021	84.55%	81.68%	83.17%	74.67%	97.57%	97.79%	89.18%
2022	90.00%	84.49%	85.06%	80.38%	97.69%	97.56%	91.60%
2023	84.98%	78.58%	81.10%	74.71%	96.89%	95.83%	90.08%
2024	82.08%	79.90%	77.81%	79.38%	97.19%	96.88%	88.33%

Table 21: DeepSeek-V3 performance on RAG task (problem generated by GPT) , Null Output is counted as 0.25.

Year	Direct Ask	RAG	RAG (GPT)	RAG (Gem)	Full (Original)	Full (GPT)	Full (Gem)
2020	83.62%	78.16%	79.02%	73.28%	94.76%	95.11%	85.13%
2021	61.92%	75.17%	76.77%	69.43%	95.81%	94.81%	84.16%
2022	84.48%	81.74%	81.49%	70.36%	96.18%	94.91%	86.83%
2023	70.83%	78.17%	79.46%	70.01%	95.72%	94.66%	85.56%
2024	75.83%	78.23%	75.10%	73.75%	95.62%	93.75%	86.35%

Table 22: DeepSeek-V3 performance on RAG task (problem generated by GPT) , Null Output is counted as 0.25.

Example 1 - Keyword Replacement

Title: NASA says object that hit Florida home is from International Space Station[a]

Question: On which date did NASA release a pallet containing old nickel–hydride batteries from the International Space Station?
A) March 8, 2021 **B)** March 11, 2021 **C)** April 22, 2024 **D)** March 8, 2020

Original Context: ... A pallet containing old nickel–hydride batteries was released from the ISS on **March 11, 2021**, after new batteries were installed. ...

LLM Revised Context: ... The debris, part of a 5,800-lb cargo pallet released from the ISS in **March 2021**, unexpectedly survived atmospheric re-entry. ...

[a] `https://en.wikinews.org/wiki/NASA_says_object_that_hit_Florida_home_is_from_International_Space_Station`

Figure 27: The news revised by LLMs omits key information about the specific date NASA released the pallet, causing the RAG system unable to determine the correct date and ultimately selecting **A**.

Example 2 - Keyword Replacement

Title: Latin American expedition of Viktor Pinchuk: meeting with the traveler took place in Yalta[a]

Question: What hobby involves collecting recordings of ethnic performers and is practiced by Viktor Pinchuk?
A) Philophony **B)** Ethnomusicology **C)** Hobo tourism **D)** Cultural preservation

Original Context: ...From every trip or an expedition, Viktor Pinchuk brings CDs with recordings of ethnic performers; the traveler's collection has already accumulated a significant number of them (not counting several hundred digital editions of world-famous musicians). The hobby is called **"philophony"**, and the subject of it is called a philophonist. ...

LLM Revised Context: ...Pinchuk, a self-described **"philophonist,"** has amassed hundreds of CDs and digital recordings of ethnic and world music. ...

[a] `https://en.wikinews.org/wiki/Latin_American_expedition_of_Viktor_Pinchuk:_meeting_with_the_traveler_took_place_in_Yalta`

Figure 28: The RAG system mistakenly selects B when using the LLM-revised text because the revision omits key details, such as the explicit mention of the hobby's name, "philophony."

Example 3 - Keyword Replacement

Title: New Zealand Navy ship HMNZS Manawanui capsizes one nautical mile from shore[a]

Question: What was the name of the Royal New Zealand Air Force aircraft that assisted in the evacuation of the crew from HMNZS Manawanui?
A) Boeing P-8 Poseidon **B)** Airbus A320 **C)** Lockheed Martin C-130J **D)** Boeing 737,

Original Context: ...They were rescued with assistance from the Rescue Coordination Centre (RCCNZ) and **a Royal New Zealand Airforce Boeing P-8 Poseidon**. ...

LLM Revised Context: ...All 75 crew were safely evacuated with assistance from the Rescue Coordination Centre and the **Royal New Zealand Air Force**.

[a] `https://en.wikinews.org/wiki/New_Zealand_Navy_ship_HMNZS_Manawanui_capsizes_one_nautical_mile_from_shore`

Figure 29: LLMs omit key information, such as the aircraft's name.

Example 4 - Abbreviation Ambiguity Misleading

Title: At least 20 die in Odesa in Russian missile strike, Ukraine reports[a]

Question: How many employees of the State Emergency Service of Ukraine were reported as victims of the missile strikes in Odesa?
A) One **B)** Five **C)** Seven **D)** Ten

Original Context: ... Among the dead are an employee of the State Service of Ukraine for Emergency Situations (SSES) and a paramedic. ... Among the victims are **seven employees of the State Emergency Service.** ...

LLM Revised Context: ... Among the deceased are a staff member of the State Service of Ukraine for Emergency Situations (SSES) and a paramedic. ... **Seven SSES personnel** were among the injured, and medical workers also sustained injuries. ...

[a] https://en.wikinews.org/wiki/At_least_20_die_in_Odesa_in_Russian_missile_strike,_Ukraine_reports

Figure 30: The original text use the full name "*seven employees of the State Emergency Service*", allowing the RAG system to correctly select C. However, the LLM's revised text abbreviated this to "*seven SSES personnel*", causing the RAG system to incorrectly choose **A**.

Example 5 - Introduction of Modifiers

Title: Arizona bans abortion for genetic abnormalities[a]

Question: What does Senate Bill 1457 in Arizona classify as a Class 6 felony?
A) Seeking or performing an abortion because of a severe fetal abnormality
B) Seeking or performing an abortion due to the presence of a genetic abnormality in the child
C) Distributing abortion-inducing drugs via courier
D) Soliciting funds for an abortion

Original Context: ... The bill makes it a Class 6 felony, the least severe, to seek or perform an abortion **"because of a genetic abnormality of the child"**, defined as "the presence or presumed presence of an abnormal gene expression in an unborn child", but not a "severe fetal abnormality" considered "incompatible with life." ...

LLM Revised Context: ... Arizona Governor Doug Ducey signed Senate Bill 1457 into law on Tuesday, effectively **banning abortions sought solely due to fetal genetic abnormalities**. The bill, which passed the Republican-controlled legislature after twice stalling and undergoing amendments to secure necessary votes, classifies seeking or performing such abortions as a Class 6 felony. ...

[a] https://en.wikinews.org/wiki/Arizona_bans_abortion_for_genetic_abnormalities

Figure 31: Although both the original and revised text explicitly excludes *"severe fetal abnormalities"*, the revised text change *"genetic abnormality"* to *"fetal genetic abnormalities"*, which leads LLMs to misinterpret the information. As a result, LLMs mistakenly select **A** based on the revised text.